软件学报 ISSN 1000-9825, CODEN RUXUEW
[doi: 10.13328/j.cnki.jos.007707] [CSTR: 32375.14.jos.007707]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

不可学习样本生成与净化方法综述*

孟若涵[1,3], 卞新玉[2], 张丛阳[2], 朱浩天[2], 付章杰[1,3]

[1](南京信息工程大学 网络空间安全学院, 江苏 南京 210044))

[2](南京信息工程大学 计算机学院、软件学院, 江苏 南京 210044)

[3](数字取证教育部工程研究中心 (南京信息工程大学), 江苏 南京 210044)

通信作者: 付章杰, E-mail: fzj@nuist.edu.cn

摘　要: 大规模深度学习模型的迅速迭代对高质量训练数据的需求不断攀升, 而这一需求通常依赖于从互联网收集海量数据, 其中不乏涉及个人图像与文本的私有信息, 因而引发潜在的隐私泄漏与伦理风险. 为缓解此类问题, 研究者提出了不可学习样本这一主动数据隐私保护策略. 该方法通过向训练数据中嵌入难以察觉的微小扰动, 使未授权模型在表面上获得较高训练精度, 却因学习到无意义的“捷径特征”而失去泛化能力. 对不可学习样本的研究进展进行系统梳理, 将现有工作划分为两大方向: 一类聚焦于不可学习样本的生成方法, 另一类则关注其净化与检测机制. 在此基础上, 进一步探讨未来的研究挑战与发展路径, 旨在为隐私数据保护提供理论支撑与技术突破.

关键词: 不可学习样本; 数据隐私保护; 数据投毒; 深度学习; 捷径学习

中图法分类号: TP18

中文引用格式: 孟若涵, 卞新玉, 张丛阳, 朱浩天, 付章杰. 不可学习样本生成与净化方法综述. 软件学报. http://www.jos.org.cn/1000-9825/7707.htm

英文引用格式: Meng RH, Bian XY, Zhang CY, Zhu HT, Fu ZJ. Review of Generation and Purification Methods for Unlearnable Examples. Ruan Jian Xue Bao/Journal of Software (in Chinese). http://www.jos.org.cn/1000-9825/7707.htm

Review of Generation and Purification Methods for Unlearnable Examples

MENG Ruo-Han[1,3], BIAN Xin-Yu[2], ZHANG Cong-Yang[2], ZHU Hao-Tian[2], FU Zhang-Jie[1,3]

[1](School of Cyber Science and Engineering, Nanjing University of Information Science & Technology, Nanjing 210044, China)

[2](School of Computer Science & School of Software, Nanjing University of Information Science & Technology, Nanjing 210044, China)

[3](Engineering Research Center of Digital Forensics (Nanjing University of Information Science and Technology), Ministry of Education, Nanjing 210044, China)

Abstract: The rapid iteration of large-scale deep learning models leads to an ever-growing demand for high-quality training data. This demand is typically met by collecting massive amounts of data from the Internet, which often include private information such as personal images and text, thus creating risks related to privacy leakage and ethical concerns. To mitigate such issues, researchers have proposed unlearnable examples as a proactive strategy for data privacy protection. This approach embeds imperceptible perturbations into training data, making unauthorized models appear to achieve high training accuracy while actually losing generalization ability due to learning meaningless “shortcut features.” This study provides a systematic review of recent advances in unlearnable examples, categorizing existing work into two main areas: generation methods and purification and detection mechanisms. Building on this foundation, future research challenges and potential directions are further explored, with the aim of providing theoretical support and facilitating technical breakthroughs in private data protection.

Key words: unlearnable example (UE); data privacy protection; data poisoning; deep learning; shortcut learning

---

* 基金项目: 国家自然科学基金 (62502221, U22B2062, 62172232); 江苏省自然科学基金 (BK20250737); 江苏省科技重大专项 (BG2024042); 江苏省高等学校自然科学研究基金 (24KJB520019); 南京市科技重大专项 (202405002); 南京市留学归国人员科技创新计划 (R2024LZ04); 南京信息工程大学人才启动经费 (2025r069)
收稿时间: 2025-09-26; 修改时间: 2026-04-27; 采用时间: 2026-05-21; jos 在线出版时间: 2026-08-19

人工智能技术的快速发展已深刻变革了诸多行业, 其核心推动力源于对海量训练数据的获取与利用. 例如, ChatGPT4 需要 13 万亿 token 的训练数据, DeepSeek 训练数据达 14.8 万亿 token. 这些数据往往不可避免地依赖网络数据抓取[1]. 然而, 部分企业机构在未获授权的情况下, 违规采集包括社交媒体、论坛评论等海量公开数据[2], 并用于商业模型训练[1,3]. 此现象不仅存在侵犯个人隐私的风险, 还可能被用于监控、商业剥削等有违伦理问题引发的活动, 已引起全球监管机构的高度关注.

为应对深度学习模型带来的数据隐私与安全问题, 现有研究从两个角度展开探索, 并都以损害模型的可用性为目标[4,5]. 一方面, 从数据访问权限控制的角度, 安全多方计算 (secure multiparty computation, SMPC) 与同态加密 (homomorphic encryption, HE) 通过对数据进行加密, 实现多方协同计算过程中的隐私保护, 防止信息泄露. 与此不同, 差分隐私 (differential privacy, DP) 采用基于扰动的数据隐私保护机制, 在数据查询或其结果中引入精确控制的噪声, 从而有效掩盖任意单个个体对最终输出的具体贡献, 实现隐私保护[6]. 另一方面, 从深度学习模型的学习过程切入, 近年来对抗样本攻击和投毒攻击揭示了训练数据层面的潜在风险: 前者通过在输入样本中注入细微扰动以误导预训练模型输出, 后者则在训练阶段混入恶意样本以干扰模型收敛. 在此背景下, 不可学习样本 (unlearnable example, UE) 技术应运而生[7], 成为防止未授权模型滥用训练数据、保护数据隐私的重要研究方向之一. 其核心思想是在训练数据中嵌入人眼难以察觉的扰动, 诱导模型学习扰动所带来的“伪特征”而非语义真实内容. 尽管模型在训练阶段可能表现出高准确率, 但其泛化能力将显著下降, 从而有效阻断对原始数据特征的提取与滥用, 实现数据隐私保护与数据正常使用价值之间的平衡.

本文将不可学习样本的相关研究归纳为两个主要类别: 不可学习样本生成方法以及相应的净化与检测方法. 遵循这一分类, 本文第 1 节概述不可学习样本的基本概念和原理. 第 2 节详细分析基于优化和基于启发式的两类不可学习样本生成方法及其评价体系. 第 3 节对基于模型训练和基于数据预处理的两类不可学习样本净化方法以及检测方法进行介绍与比较. 第 4 节对不可学习样本生成与净化方法的未来发展方向进行展望. 本文具体框架可见表 1.

表 1　不可学习样本生成与净化方法综述框架

大类	子类	次子类	简介
不可学习样本概述	—	—	基本概念以及原理
不可学习样本生成方法	基于优化的不可学习样本生成方法	直接梯度优化方法	直接对扰动生成目标函数使用梯度下降策略优化
		基于特征空间优化方法	在特征空间上干扰原始数据分布
		基于不同目标优化方法	不同任务、模态数据下的优化方法
	基于启发式的不可学习样本生成方法	—	引入偏差信息, 无需替代模型生成不可学习样本
	不可学习样本生成方法评价体系	—	介绍评价不可学习样本生成方法5个主要方面, 以及创新评价指标和框架
不可学习样本净化与检测方法	基于模型训练的不可学习样本净化方法	—	直接使用不可学习样本训练模型, 仅修改训练方案
	基于数据预处理的不可学习样本净化方法	基于简单图像变换的净化方法	对不可学习样本先进行压缩、非线性变换等操作, 再训练模型
		基于图像重生成的净化方法	使用生成模型使用扩散模型重生成图像
		其他净化方法	正交投影方法、针对卷积不可学习样本的净化方法
	不可学习样本检测方法	—	检测收集到的数据是否为不可学习样本
不可学习样本生成与净化方法的未来发展方向	不可学习样本生成方法未来发展方向	—	未来不可学习样本生成方法的发展需提升其在复杂防御下的鲁棒性、跨任务与模态的通用性, 并优化计算效率与解决部分数据保护的局限性
	不可学习样本净化方法未来发展方向	—	未来不可学习样本净化方法的发展需提升其通用性, 优化净化效率, 降低数据依赖, 并将净化方法扩展到除图像以外的其他模态

1 不可学习样本概述

互联网上公开可用的个人数据常在未经授权的情况下被爬取用于商业模型训练, 这不仅会引发严重的隐私侵犯问题, 还涉及知识产权保护等法律争端. 传统的对抗攻击 (adversarial attack) 通过在测试阶段添加微小扰动到输入样本中来误导模型预测, 主要影响的是模型在推理阶段的性能[4]; 而后门攻击 (backdoor attack) 则通过在训练数据中植入特定触发器来操控模型行为, 需要攻击者能够控制训练过程[5]. 这两种攻击方式虽然都能影响模型性能, 但都无法从根本上防止数据被用于模型训练.

与传统攻击方法不同, 不可学习样本代表了一种新型的数据投毒范式, 由 Huang 等人[7]提出. 其核心目标是在训练数据中注入难以察觉的特定扰动, 以干扰模型训练过程中的梯度传播, 使参数更新朝向误导性方向快速收敛至局部极值. 这种人为设计的收敛特性会使模型在训练阶段呈现出极高的准确率, 实际上却未能学习到有效的语义特征, 从根本上破坏模型从训练数据中学习有效特征的能力. 不可学习样本技术的关键创新在于, 其干扰操作发生在数据发布前, 先验性地对原始数据进行扰动处理, 从而阻断任何基于该数据训练的模型在性能上的有效提升. 与传统对抗样本攻击主要作用于模型推理阶段不同, 不可学习样本直接作用于训练阶段, 从根源上破坏模型对有用特征的学习能力; 而相较于后门攻击需控制模型训练流程, 不可学习样本仅需对数据进行一次性预处理, 具有更强的实用性与可部署性. 从实现机制上看, 不可学习样本通过注入精心设计的扰动, 使模型在训练中趋向学习扰动模式中的“捷径特征”, 而非数据本身所蕴含的真实语义特征, 其生成流程框架如图 1 所示. 从隐私保护角度出发, 不可学习样本提供了一种主动防御的数据安全方案. 相较于传统如差分隐私等被动保护机制[6], 不可学习样本直接在数据层面预先构建扰动屏障, 即使数据被非法收集, 也能有效阻止其被深度模型利用, 保障数据安全性. 这种机制尤其适用于需要公开共享数据但又需防止未授权滥用的场景, 如医疗数据共享、人脸识别数据保护等领域.

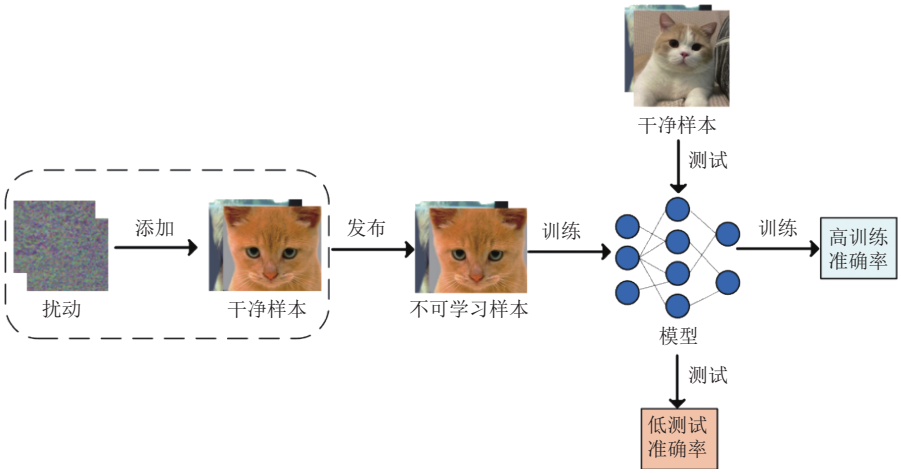

图 1 不可学习样本生成流程

本文从两个视角对不可学习样本的相关研究进行系统性梳理: 一是不可学习样本生成方法, 二是不可学习样本净化与检测方法. 在生成方法方面, 本文将其进一步划分为基于优化和基于启发式的不可学习样本生成方法以及相关评价体系. 基于优化的方法通过数学建模将不可学习样本的生成问题转化为目标函数优化问题, 其核心在于构造扰动, 系统性地影响模型训练中的特征提取和梯度更新过程. 此类方法又可细分为直接梯度优化、特征空间优化和不同目标优化等次子类. 直接梯度优化通过替代模型梯度下降策略优化扰动, Huang 等人[7]提出误差最小化 (error minimizing, EM) 方法来构造不可学习样本. 通过优化扰动使模型训练时的梯度信号被误导, 导致模型快速收敛至局部极值, 无法学习真实特征. 在此基础上, 对抗性诱导噪声 (adversarially inducing noise, ADVIN)[8]、灰度化增强 (grayscale image augmentation, GrayAugs)[9]等进一步增强了扰动的鲁棒性.

特征空间优化方法则通过操纵数据在特征空间的分布来影响数据的可学习性, 如特征混淆 (entangled feature,

EntF)[10]和可转移不可学习样本 (transferable unlearnable example, TUE)[11]等. 针对不同任务和模态, 研究者还提出不同目标优化, 包括分类任务中的稀疏感知局部掩码 (sparsity-aware local masking, SALM)[12]、生成任务中的增强不可学习扩散扰动 (enhanced unlearnable diffusion perturbation, EUDP)[13]、分割任务中的分割不可学习 (unlearnable segmentation, UnSeg)[14]等. 基于启发式的方法则通过引入特定的偏差信息, 使模型在训练期间寻找“捷径”而忽视数据的固有特征. 这类方法无需依赖替代模型, 计算效率较高. 包括单像素捷径 (one pixel shortcut, OPS)[15]、自回归 (autoregressive, AR)[16]、基于卷积的不可学习数据集 (convolution-based unlearnable dataset, CUDA)[17]等. 不可学习样本的评价体系涵盖有效性、不可感知性、鲁棒性等. 为弥补传统评价体系的不足, 最新研究提出了更加系统的评价体系. 可用性中毒攻击基准 (availability poisoning attacks benchmark, APBench)[18]建立了标准化的评价体系. Ye 等人[19]从模型优化角度出发, 提出了新的创新评价指标.

在净化与检测方法方面, 本文将其划分为基于模型训练和基于数据预处理的方法. 基于模型训练的方法通过调整训练方案或优化损失函数, 使模型能够从不可学习样本中提取出有用的泛化特征, 从而恢复模型在干净测试集上的性能. 这类方法的核心在于削弱不可学习样本对模型训练的负面影响, 同时增强模型对干净数据的适应能力, 包括对抗训练 (adversarial training, AT)[10]、对抗增强 (adversarial augmentation, AA)[11]等. 基于数据预处理的方法在模型训练之前对数据进行处理, 目的是破坏或去除不可学习样本中的扰动, 恢复数据的可学习性. 这类方法的核心在于通过预处理操作消除不可学习扰动对数据的影响, 从而使模型能够在处理后的数据上正常训练. 这类方法又可细分为简单图像变换、图像重生成等次子类. 简单图像变换通过对图像进行一些简单的操作, 以此削减图像中的噪声扰动, 方法包括图像捷径压缩 (image shortcut squeezing, ISS)[20]、针对不可学习数据集的非线性变换 (nonlinear transformations against unlearnable datasets, NTAUD)[21]等. 图像重生成使用扩散模型通过正向扩散过程将不可学习样本中的扰动逐渐去除, 然后通过反向去噪过程恢复出干净的图像, 方法包括 AVATAR[22]、可学习样本 (learnable example, LE)[23]等. 检测方法的核心在于识别数据是否受到不可学习样本的污染, 从而在模型训练之前及时采取措施. 检测方法通常利用不可学习样本的特性, 如线性可分性、模型对不可学习样本的快速适应性等, 进行判断. 方法有拓扑数据分析 (topological data analysis, TDA)[24]、迭代过滤 (iterative filtering, IF)[25]等. 后文图 2 与图 3 分别展示了不可学习样本生成方法的研究时间线图和不可学习样本净化与检测方法的研究时间线.

不可学习样本生成方法与净化方法之间存在着动态对抗关系, 生成方法迭代演进, 到净化方法突破, 再到生成方法鲁棒生成升级. 如后文图 4 所示, 我们列出来具有代表性的生成与净化方法, 展示两者之间的攻防关联. 误差最小化 (EM)[7]作为首个典型的不可学习样本生成方法, 直接推动了以对抗训练 (AT)[26]为代表的早期净化方法的出现; 为抵御 AT 的净化, 鲁棒误差最小化 (robust error-minimizing, REM) 噪声[27]通过引入对抗扰动和期望变换提升了对 AT 的抵抗能力; 数据增强类净化方法的发展, 又催生了灰度化增强 (GrayAugs)[9]等对数据增强具有鲁棒性的生成方法被提出. 随着生成方法向无代理模型的启发式生成方法演进, 基于卷积的不可学习数据集 (CUDA)[17]这种基于卷积捷径的生成方法出现, 随即被随机锐化核 (random sharpening kernel, RSK)[21]通过破坏固定特征关联实现净化; 而对抗增强 (AA)[11]、图像捷径压缩 (ISS)[20]等净化方法的出现, 进一步推动生成方法向更具语义捷径的方向演进, 如语义图像深度隐藏 (deep hiding, DH)[28]方法通过嵌入语义图像到干净图像中, 构造具有语义信息的捷径来实现对多种净化方法的鲁棒性, 但其在 AVATAR[22]净化方法下仍存在局限性.

2 不可学习样本生成方法

2.1 基于优化的不可学习样本生成方法

基于优化的方法通过数学建模将不可学习样本的生成转化为目标函数优化问题, 其核心在于构造扰动, 从系统层面干扰模型的特征提取和梯度更新过程, 进而削弱模型对真实语义信息的学习能力.

2.1.1 直接梯度优化方法

直接梯度优化方法通过基于梯度下降的优化算法对扰动进行迭代更新, 旨在使训练模型无法从受扰动的数据中提取有效特征. 扰动的优化目标是最小化模型在扰动数据上的损失, 引导模型学习扰动特征而非原始数据特征.

并通过设计针对对抗训练 (AT) 和数据增强 (data augmentation) 的防御措施, 增强扰动的鲁棒性.

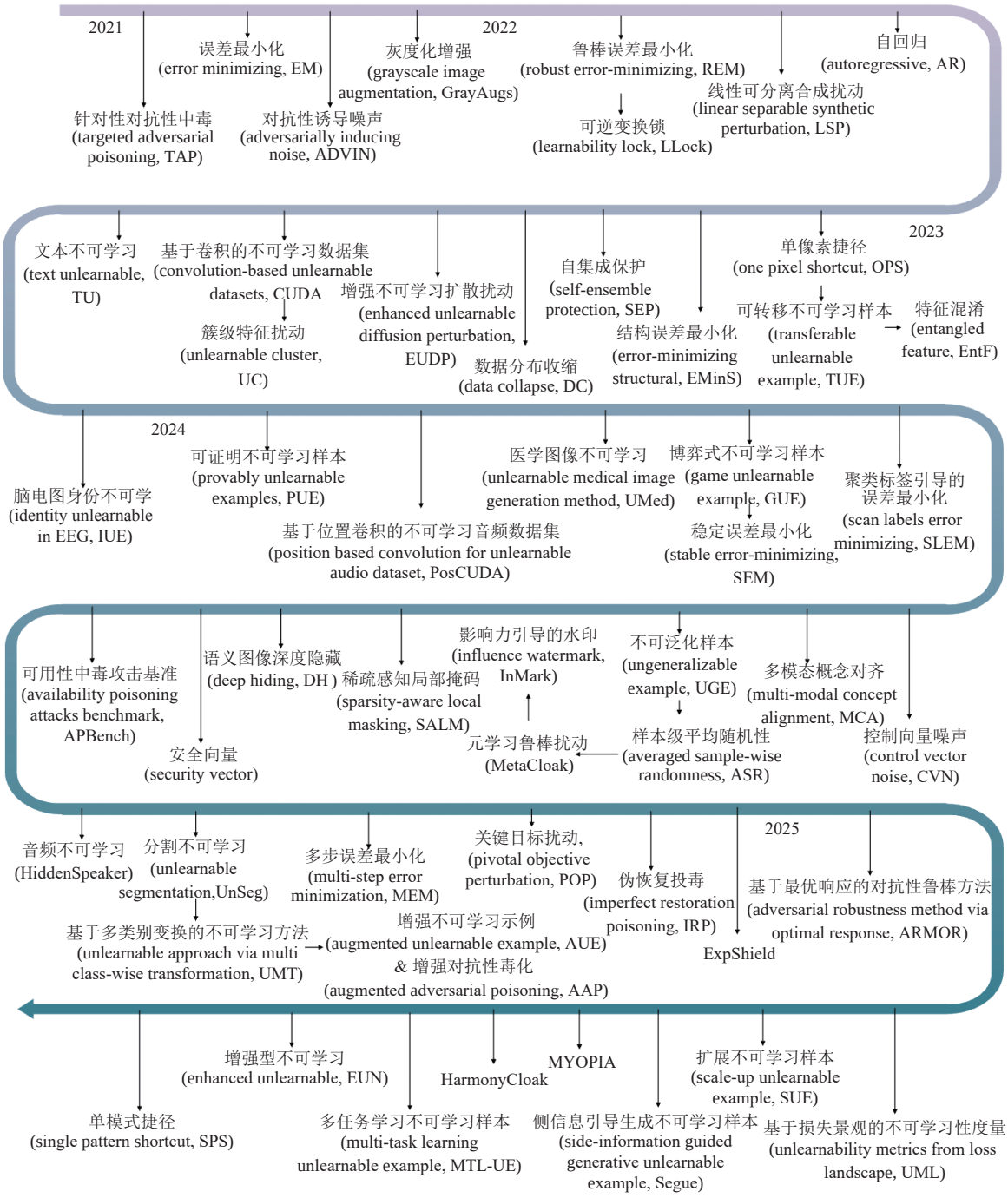

图 2　不可学习样本生成方法研究时间线

Huang 等人[7]通过构造一个双层优化问题提出了误差最小化 (EM) 方法, 用于生成不可学习样本扰动. 该方法的内层优化旨在寻找合适的噪声以最小化分类损失, 外层优化则更新模型参数以辅助扰动的有效性. 具体而言, EM 首先初始化一个替代模型用于模拟目标模型的训练行为, 并根据扰动的作用粒度 (如样本级或类别级) 确定扰

动类. 随后, 通过梯度下降优化模型参数. 误差最小化噪声生成阶段, 使用投影梯度下降算法 (projected gradient descent, PGD) 迭代优化噪声 δ 以最小化分类损失, 投影操作约束噪声幅度. 最后当模型分类误差小于阈值时停止迭代. 双层优化问题公式如公式 (1) 所示.

$$\min_{\theta}\mathbb{E}_{(x,y)\sim D_c}\left[\min_{\|\delta\|_p\leqslant\varepsilon}\mathcal{L}(f_\theta(x+\delta),y)\right] \tag{1}$$

其中, θ 表示模型参数, $\mathcal{L}$ 为分类损失函数, δ 为不可学习扰动. 内层优化旨在寻找扰动 δ, 使添加扰动后的训练样本 $x+\delta$ 分类损失最小; 外层则通过优化模型参数 θ, 进一步强化这一损失最小化过程.

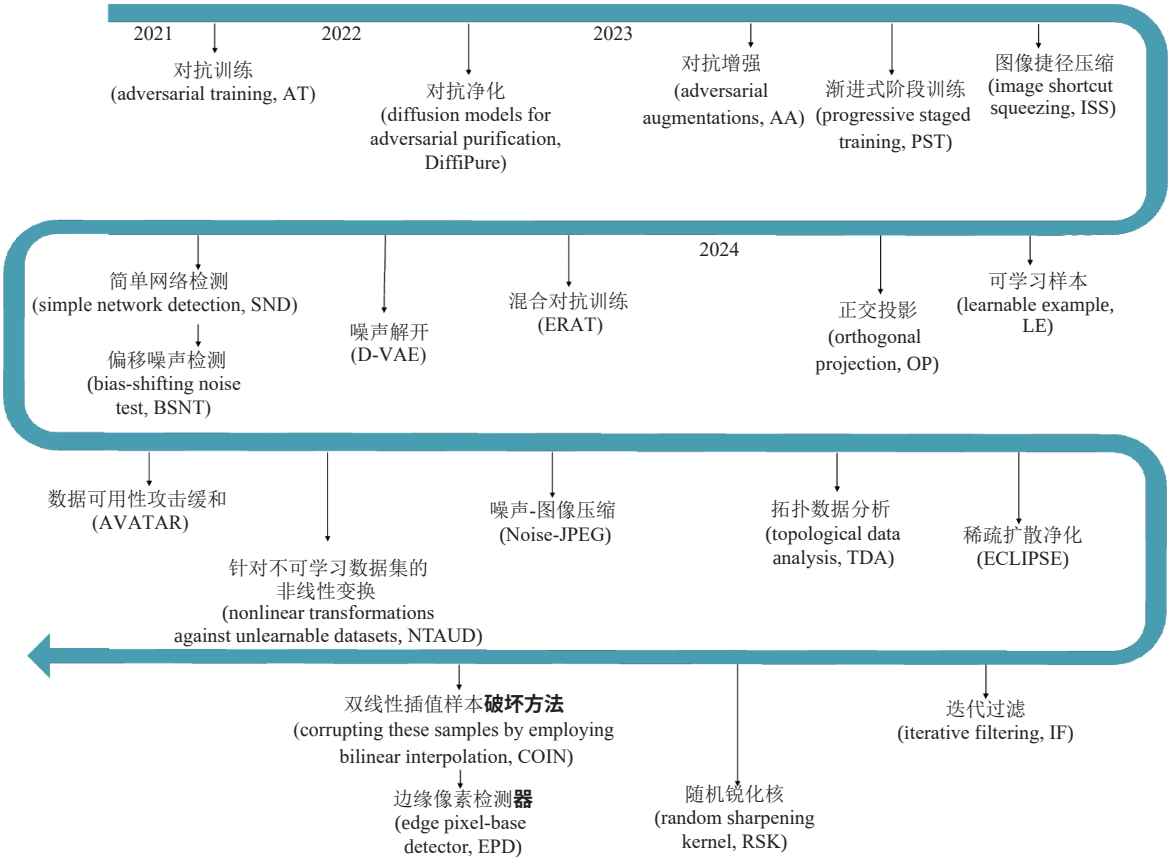

图 3 不可学习样本净化与检测方法研究时间线

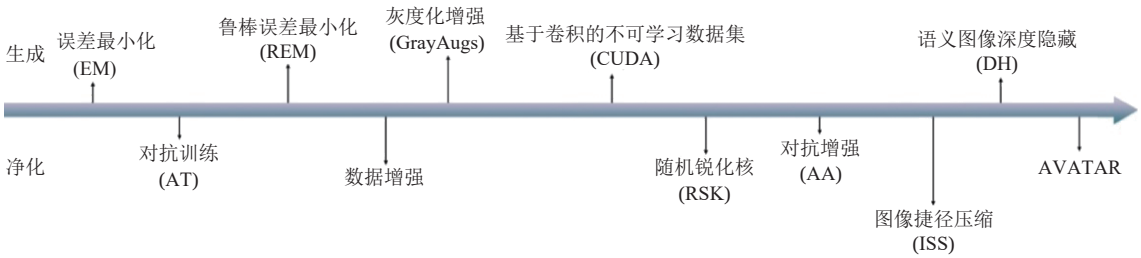

图 4 不可学习样本生成与净化方法动态对抗图

通过该优化策略, 模型在训练阶段能够对扰动后的数据“完美分类”, 却未学习到有用的语义特征, 进而导致在测试阶段的表现如同随机猜测, 实现不可学习的目标. 实验表明, EM 方法在多种模型架构和数据集上均具有效性, 且具备良好的可迁移性. 例如, 在 ImageNet 上生成的扰动可成功迁移至 CIFAR-10. 然而, 该方法在对抗性训练下抵抗性较弱. 例如, 在 CIFAR-10 上, 该方法在普通训练下可将模型测试准确率降至 19.93%, 但在对抗训练下模型测试准确率回升至 79%. 为提升对抗训练下的防御能力, Fu 等人[27]提出鲁棒误差最小化 (REM) 噪声方法, 在原有双层优化框架中引入对抗扰动和期望变换 (expectation over transformation, EOT) 机制, 生成能够抵御对抗训练的不可学习样本. 其核心思想是“以对抗训练的方式防御对抗训练”, 增强了扰动在多种训练动态下的稳定性, 但也显著提升了计算成本; 同时, 当对抗扰动半径未知时, 需设置较大的扰动范围以确保防御效果. 进一步地, Wang 等人[8]提出对抗性诱导噪声 (ADVIN) 方法, 通过引入标签诱导机制, 实现对目标模型的深度污染. 该方法采用双层循环迭代策略: 内层循环通过最小化扰动样本与误导目标标签间的损失, 更新噪声并迫使噪声分布与模型梯度下降轨迹正交. 外层循环进行对抗训练, 将对抗噪声约束在人类不可感知的 Lp 范数邻域内, 直到污染成功率 (poisoning success rate, PSR) 大于阈值时终止. 实验中, ADVIN 显著降低了对抗训练的鲁棒测试准确率, 使在 CIFAR-10 上的测试准确率从 51.71% 降至 0.57%. 但其使用的扰动强度较大 (32/255), 在一定程度上损害了扰动的不可感知性. 为进一步提升对抗训练下的防御性扰动质量, 样本级平均随机性 (averaged sample-wise randomness, ASR)[29]引入平均样本随机性约束机制, 该机制在对抗训练框架中强制模型对干净数据输出趋于均匀的分布, 从而提高扰动的不可学习性, 同时使模型梯度在对抗扰动和随机性扰动之间形成博弈关系, 从而增强模型对敌手样本的抗干扰能力.

不同于以往通过增强替代模型鲁棒性间接提升不可学习样本的稳定性, 近期研究开始关注从防御噪声本身的性质出发, 以提升扰动在更广泛预处理或变换条件下的抗干扰能力. Liu 等人[30]提出了稳定误差最小化 (stable error-minimizing, SEM) 噪声, 训练防御噪声抵抗随机扰动, 而非只针对耗时的对抗扰动. 不可学习样本容易受灰度化处理影响, Liu 等人[9]提出灰度化增强 (GrayAugs), 结合灰度化和数据增强技术, 消除 UE 中依赖颜色通道的扰动, 生成更具空间变化的扰动, 使其在灰度预处理下仍能保持有效性, 但在较大数据集上的表现有限. 张弈晟等人[31]提出一种聚类标签引导的误差最小化 (SCAN labels error minimizing, SLEM) 方法, 其核心思路是结合 SCAN 深度聚类与 EM 无监督数据投毒. 具体而言, 该方法首先利用近邻语义聚类 SCAN (semantic clustering by adopting nearest neighbor) 算法对无标签图像进行聚类并生成伪标签. 随后, 通过训练模型筛选低损失样本, 并在这些样本上应用 EM 算法生成噪声[7]; 对于高损失样本, 则借助干净模型重新标注, 并采用投影梯度下降优化噪声. 在此过程中引入期望变换, 以增强对数据增强操作的鲁棒性. 然而, 由于 SCAN 生成伪标签的准确率有限, 其整体噪声生成效果仍受到一定制约. 此外, 增强不可学习示例 (augmented unlearnable example, AUE) 和增强对抗性毒化 (augmented adversarial poisoning, AAP)[32]将对比学习中的强数据增强机制引入监督学习的攻击框架中, 使得生成的扰动能够同时干扰监督学习模型的特征学习和对比学习模型的特征对齐与分布均匀性. 最终, 该类扰动不仅削弱了监督学习模型的性能, 也能有效破坏对比学习模型的表现. 基于最优响应的对抗性鲁棒方法 (adversarial robustness method via optimal response, ARMOR)[33]通过在替代模型中引入非局部模块, 有效提升替代模型对样本全局特征的捕获能力. 同时, 该方法采用类别自适应的数据增强策略选择机制, 可根据不同类别样本的特性动态调整增强方式. 具体而言, ARMOR 通过最大化同类数据增强样本与扰动样本之间的梯度余弦相似度, 实现最优梯度对齐, 从而选择最合适的数据增强策略. 最终, 经过迭代优化过程, 该方法能够生成更具鲁棒性的扰动. 侧信息引导生成不可学习样本 (side-information guided generative unlearnable example, Segue)[34]则利用预训练生成模型产生扰动, 并引入侧信息 (如真实标签或伪标签嵌入) 以增强扰动的转移性. 同时, Segue 在生成过程中加入扭曲层等数据变换操作, 以进一步提升样本的鲁棒性.

综上所述, 直接梯度优化方法的演进清晰地体现了不可学习样本生成与净化方法之间的关系. 从早期 EM 方法针对标准训练有效, 到 REM、ADVIN 等方法通过引入期望变换, 标签诱导和随机性约束等机制提升抵御对抗训练的能力; 与此同时, GrayAugs、SLEM 等方法则针对数据增强策略进行鲁棒性优化. 这一演变路径表明, 生成方法正在不断进行鲁棒化升级, 以应对实际部署中日益复杂和多样化的净化方法. 不可学习样本生成方法实际部署对比如表 2 所示.

表 2 不可学习样本生成方法实际部署对比

方法		代表方法	计算开销	模型依赖
基于优化的生成方法	直接梯度优化	EM[7]、REM[27]	高	需代理模型
	特征空间优化	EntF[10]、TUE[11]	中	需特征提取器
	不同目标优化	EUDP[13]、UnSeg[14]	中高	需目标特定模型
基于启发式的生成方法		CUDA[17]、LSP[35]	低	无

2.1.2 基于特征空间优化方法

在不可学习样本的研究中, 特征空间优化是一类核心方法. 通过操纵数据在特征空间分布, 破坏数据的可学习性, 概略方法如图 5 所示.

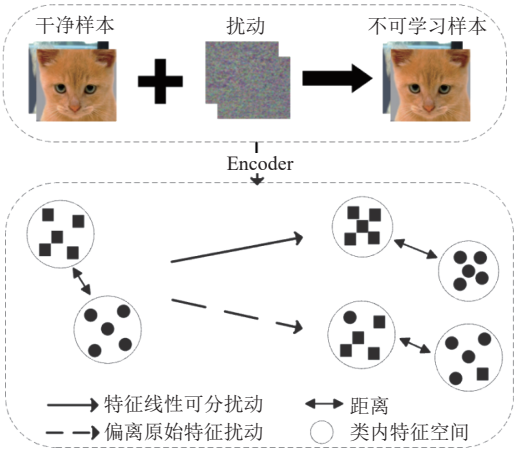

图 5 特征空间优化方法概览

基于特征空间优化方法主要分为特征线性可分扰动和偏离原始特征扰动, 其中特征线性可分扰动构造可分但错误的边界, 偏离原始特征扰动则破坏数据的可分性. Ren 等人[11]提出可转移不可学习样本 (TUE), 引入类间可分性判别方法, 最小化类内距离, 最大化类间距离, 加强扰动的线性可分性:

$$\min \frac{S_W}{S_B} \tag{2}$$

其中, S_W 代表两个不同类的类内距离, S_B 代表两个不同类的类间距离. 通过双层约束, 实现在不同训练环境和数据集上的可转移性. 然而, 这类方法往往依赖显式标签以进行类间可分性判别, 限制了其在完全无监督场景中的适用性. 针对该问题, 簇级特征扰动 (unlearnable clusters, UC)[36]利用代理模型来提取图像的特征表示, 并使用 K-means 聚类算法将样本分为多个语义簇. 随后, 为每个簇构造簇级扰动, 通过最小化其与伪造簇中心的距离损失函数, 破坏数据分布的一致性和差异性, 生成不可学习样本. 尽管 UC 在语义层面实现了结构性干扰, 但在大规模数据集上计算复杂度较高, 扰动半径过大, 影响扰动的不可感知性. 由于 UC 方法需要对每个簇都进行最小化损失函数生成簇级扰动, 需要大量的计算资源. 为提升 UC 的扩展能力, 拓展不可学习样本 (scale-up unlearnable example, SUE)[37]利用 Summit 超级计算机的计算能力来检查不同批量大小对 UC 方法[36]性能的影响, 分析不同批量大小对扰动效果的影响, 并将特征提取模块嵌入模型计算图中以提升效率. 实验发现, 批量大小与模型性能间的关联性高度依赖于具体数据集结构, 说明可扩展性的优化仍需任务感知设计. 针对数据授权使用与非授权模型访问控制问题, 不可泛化样本 (ungeneralizable example, UGE)[38]能够实现数据的授权合法使用和防止非授权网络的访问. 非授权模型与可转移不可学习样本 (TUE)[11]概念类似, 但不是直接优化类别的统计距离, 而是通过三元损失与多模态预训练模型 (contrastive language-image pre-training, CLIP) 实现, 计算在共享特征空间的特征距离损失, 最大化扰动样本与原始样本的图像特征差异, 将扰动样本拉向正确标签对应的文本特征, 远离最不相关的负样本对应的文本特征,

阻止非授权网络的数据访问. 在深层语义干扰方向, 语义图像深度隐藏 (DH)[28]利用语义图像生成模型, 基于 CLIP 文本提取器对所有文本提示进行 K-means 聚类, 将文本提示和边缘地图作为输入, 生成具有隐含语义图像, 使用可逆神经网络 (invertible neural network, INN) 将预定义的语义图像自适应地隐藏到干净图像中, 并操纵隐特征来规范化扰动的类内方差. Fang 等人[39]提出数据分布收缩 (data collapse, DC) 方法, 通过梯度匹配技术估计数据分布并引导同类样本在特征空间聚集, 同时结合对抗训练优化噪声生成过程, 有效破坏数据可学习性的同时抵抗了对抗攻击, 但需要对抗训练和额外的梯度估计器训练, 扩展到大尺度数据集时效率较低. 此外, Wen 等人[10]考虑合理设置的情况, 即防御噪声小于等于对抗训练噪声的设置下, 设计特征混淆 (EntF), 使不同类别的样本在特征空间中重叠破坏特征的可分性, 即将每个样本的特征表示尽可能远离其原始类别的特征中心, 从而破坏类内紧凑性. EntF 分为两个部分, 第 1 部分将特征推离原始类别中心 c_y:

$$\max_{\delta}\left\|F(x)-c_y\right\|_2 \tag{3}$$

第 2 部分使特征靠近相邻的其他类别 $c_{y'}$:

$$\min_{\delta}\left\|F(x)-c_{y'}\right\|_2 \tag{4}$$

在此基础上, 利用 KL 散度约束使合法特征与对抗特征在隐空间形成拓扑纠缠结构, 从而在模型训练时同步激活正常模式与对抗模式的竞争性梯度信号. 实验表明, EntF 在合理设置下能够有效降低模型测试准确率. Chen 等人[40]使用多模态概念对齐 (multi-modal concept alignment, MCA), 提出一种基于多模态对齐的通用扰动生成器. 该方法通过利用 CLIP 的语义对齐能力, 在共享嵌入空间中将图像表征与相反语义的文本概念距离最小化, 同时使其与相似语义的文本概念距离最大化, 从而实现对图像分类结果的定向干扰. 由于该方法不依赖显式的类别标签关联, 因此生成的扰动具有跨数据集的直接可迁移性. 然而, 受限于 CLIP 模型对高质量特征提取的要求, 该方法在低分辨率图像上的应用效果存在明显局限.

2.1.3 基于不同目标优化方法

(1) 任务特异性方法

本文从分类、生成、分割及多任务等不同任务类型出发, 根据任务特异性对不可学习样本方法进行归类, 具体见表 3.

表 3 任务特异性方法分类

任务类型	子类	方法名称	简介	优点	缺点
分类任务	—	稀疏感知局部掩码 (SALM)[12]	对医疗图像关键区域进行扰动	1) 计算高效 2) 对医疗图像保护效果好	不同数据集效果依赖扰动像素比例
		博弈式不可学习样本 (GUE)[41]	Stackelberg博弈框架和BOME一阶优化算法训练扰动生成器	1) 理论保障 2) 强泛化性	较大对抗半径下效果下降
		可证明不可学习样本 (PUE)[42]	可学习性认证机制强化扰动的鲁棒性	理论保障	参数空间限制
		可逆变换锁 (LLock)[43]	可逆变换生成扰动	具有可逆性	对抗训练下授权用户可学习性未知是否可恢复
		针对性对抗投毒 (TAP)[44]	对抗样本作为不可学习样本	实现简单	对抗训练下效果显著下降
		自集成保护 (SEP)[45]	利用代理模型中间检查点构成自集成, 并结合特征对齐与随机方差缩减生成扰动	具有可转移性	对抗训练下效果显著下降
生成任务	图像生成	增强不可学习扩散扰动 (EUDP)[13]	优先扰动关键的时间步	动态时间步采样	计算成本高
		元学习鲁棒扰动 (MetaCloak)[46]	元学习与期望变换增强扰动	1) 具有可转移性 2) 具有鲁棒性	计算成本高
		MYOPIA[47]	高斯分布采样时间戳, 恰可察觉差异约束扰动范围, 期望变化增强扰动	具有鲁棒性	计算成本高

表 3 任务特异性方法分类 (续)

任务类型	子类	方法名称	简介	优点	缺点
生成任务	文本生成	安全向量 (security vector)[48]	参数高效调优思想, 有害行为的学习与主干参数分离	具有高效性, 仅需100条有害样本训练安全向量	高学习率下失效
		ExpShield[49]	干扰模型的token嵌入	具有高效性	极端对抗场景受限
	音频生成	关键目标扰动 (POP)[50]	重构损失作为扰动优化目标, 实施位置固定的局部扰动策略	1) 计算成本较低 2) 具有可转移性	对抗性攻击的局限性
		HarmonyCloak[51]	利用频率掩蔽技术和窗口策略, 将噪声隐藏在音乐的掩蔽区域内	支持多种音乐格式	未涵盖人声的复杂特性
分割任务	—	分割不可学习 (UnSeg)[14]	可学习噪声注入, 动态噪声强度调整	对抗训练下具有鲁棒性	1) 对简单形状和纹理的保护较弱 2) JPEG压缩下抵抗性较弱
		医学图像不可学习 (UMed)[52]	协同注入轮廓与纹理感知扰动	针对性保护性能好	计算复杂度较高
多任务	—	基于多类别变换的不可学习方法 (UMT)[53]	差异化的几何变换扰动	1) 对分类、分割任务均有效 2) 具有可逆性	无法应用于变换不变的未授权网络
		多任务学习不可学习样本 (MTL-UE)[54]	任务标签嵌入与嵌入正则化优化	在多任务数据集上有效	任务数量增加, 效果下降
		脑电图身份不可学习 (IUE)[55]	最小化任务分类的均方误差和身份分类的交叉熵	有效保护身份隐私的同时保持正常任务分类	小样本下效果有限
		单模式捷径 (SPS)[56]	向目标边界框内嵌入特定的中频模式	首次实现针对目标检测的不可学习样本	对抗训练下抵抗性较弱

(a) 分类任务

第 2.1.1 与 2.1.2 节介绍的直接梯度优化和特征空间优化方法, 均针对分类任务实现了不可学习样本的生成. 然而, 在分类任务中还有其他具有代表性的不可学习样本生成策略.

稀疏感知局部掩码 (SALM)[12]针对医疗图像的稀疏特性, 通过基于梯度排序评估像素重要性, 选择关键区域施加扰动, 从而缩小搜索空间并提升保护效率. Liu 等人[41]提出博弈式不可学习样本 (game unlearnable example, GUE), 构建了一个由“毒物攻击者”与“分类器受害者”组成的统一博弈框架. 在该框架下, 利用基于编码器-解码器的生成器为每个样本标签对添加细微扰动, 以破坏分类器性能. 最终通过 BOME 一阶优化算法求解双层优化问题, 得到能在多种场景下有效防御的不可学习样本, 但在较大对抗半径下保护效果会失效. Wang 等人[42]提出可证明不可学习样本 (provably unlearnable examples, PUE), 引入 (q,η) 可学习性的概念, 用以量化未经授权模型在受保护数据集上可达到的干净测试性能上限. 该方法通过参数平滑限制模型性能, 并在替代模型训练过程中结合随机权重扰动, 进一步降低 (q,η) 可学习性, 从而实现更强鲁棒性的数据保护. Peng 等人[43]提出可逆变换锁 (learnability lock, LLock), 基于线性或卷积可逆变换生成扰动, 并通过双层优化调整扰动强度. 该方法允许授权用户通过密钥恢复数据的可学习性, 从而实现特定数据集的可学习性控制. 但在对抗训练场景下, 密钥能否有效恢复仍未得到验证. Fowl 等人[44]提出针对性对抗投毒 (targeted adversarial poisoning, TAP), 直接利用对抗样本作为不可学习样本. 为提升扰动的可转移性, Chen 等人[45]进一步提出自集成保护 (self-ensemble protection, SEP) 方法. 该方法在代理模型训练中保存多个中间检查点, 并在扰动生成过程中, 利用特征对齐损失计算各检查点的梯度, 再通过平均与随机方差缩减技术稳定化集成梯度, 最终迭代更新扰动. SEP 能够有效削弱多种模型的训练性能, 但其在对抗训练下的保护能力仍显著下降.

(b) 生成任务

文本到图像生成模型通常能够基于少量参考照片生成高度个性化的图像. 然而, 当其被滥用于生成误导性或恶意内容时, 可能会对个人隐私与安全构成严重威胁. 为此, 针对图像生成模型的扰动设计, 其核心目标是最小化

重构损失, 其形式化表达如下:

$$\|\epsilon-\epsilon_\theta(x_{t+1}+\delta,t,c)\|_2^2 \tag{5}$$

其中, ϵ 表示在扩散过程中第 t 个时间步添加干净图像的真实噪声, ϵ_θ 为模型在同一时间步对噪声的预测结果. 不可学习样本的生成目标, 即在输入中注入适当扰动, 使真实噪声与预测噪声之间的差异最小化, 从而削弱模型在训练过程中的可学习性. Zhao 等人[13]提出增强不可学习扩散扰动 (EUDP), 通过双层最大-最小优化与时间步感知扰动设计, 实现对扩散模型训练过程的精准破坏. 外层优化目标是最大化训练损失, 内层则最小化模型参数更新. 同时考虑不同时间步 t 在扩散训练中的重要性差异, EUDP 结合噪声调度机制, 优先对关键时间步施加扰动, 从而进一步提升保护效果. 实验表明, EUDP 显著降低了生成图像质量, 并在风格迁移与文本到图像任务中展现出有效性. 但该方法涉及双层优化迭代、多时间步耦合及大规模参数计算, 计算成本较高. 元学习鲁棒扰动 (MetaCloak)[46]则基于元学习与期望变换生成扰动. 该方法构建覆盖不同训练阶段的替代模型池, 并在元学习框架下产生具有迁移性的模型无关扰动. 同时结合数据变换鲁棒性优化, 有效增强了对扩散模型个性化生成能力的干扰. 然而, 在依赖高频细节重建的超分辨率任务中, 由于扰动频率失配, 其保护效果受到限制. MYOPIA[47]采用类似 EM 的双层优化方法[7], 鉴于扩散模型中缺乏标签, 该方法通过最小化预测噪声与计划噪声之间的 Lp 范数距离来优化扰动, 并基于高斯分布采样时间步以提升干扰效果. 同时利用恰可察觉差异 (just-noticeable-difference, JND) 约束, 将扰动限制在人眼不敏感的区域, 并结合期望变换增强扰动在预处理操作下的鲁棒性.

在文本生成方面, Zhou 等人[48]为防止大语言模型在微调过程中学习有害内容, 提出安全向量 (security vectors) 微调. 该方法利用 EM 误差最小化的方法[7]训练“安全向量”, 使模型在处理有害数据时输出预设的“已学习”响应, 从而避免进一步优化有害方向. 在微调阶段, 安全向量与主干参数分离, 以保持模型整体性能; 在推理阶段, 禁用安全向量, 确保模型输出无害内容. ExpShield[49]旨在通过扰动干扰大语言模型的 token 嵌入与训练过程. 其方法分阶段优化: 初始阶段采用随机扰动, 在文本序列中随机嵌入噪声 token; 进一步阶段利用代理模型计算条件预测概率, 并对低概率 token 定向插入扰动; 同时设计基于特殊 token (如异常 token 与 OOV token) 的干扰机制, 以持续增强防御效果.

在语音生成方面, 关键目标扰动 (pivotal objective perturbation, POP)[50]针对语音合成模型提出对抗扰动生成方法, 以优化重构损失为核心. POP 根据模型架构采取差异化策略: 对窗口化生成器, 采用固定局部扰动以降低复杂度并保持隐蔽性; 对轻量级模型, 则采用全局扰动以实现全面防护. 实验结果显示, 该方法在保持扰动隐蔽性的同时显著提升了语音不可用性指标, 词错误率由 21.94% 提高至 127.31%. HarmonyCloak[51]则面向音乐生成, 提出基于双层优化的防护策略. 该方法结合频率掩蔽技术与窗口策略, 将噪声隐藏在音乐的掩蔽区域中, 从而在保证音质和防止风格模仿的同时实现有效防护. 目前, HarmonyCloak 主要针对乐器场景, 尚未扩展至包含人声的复杂生成任务中.

(c) 分割任务

Sun 等人[14]提出了分割不可学习 (UnSeg), 利用预训练的图像分割模型 (segment anything model, SAM) 作为噪声生成器, 在其掩码解码器中引入可学习的噪声令牌, 并通过自注意力与交叉注意力机制生成扰动. 该方法采用双层最小化优化框架, 交替训练噪声生成器与替代模型, 并优化噪声以最小化替代模型的训练误差. 同时通过比例缩放噪声强度 ϵ 来稳定训练过程, 从而确保在推理阶段噪声仍能保持足够强度. 由于 JPEG 压缩会舍弃高频信息, UnSeg 对 JPEG 的鲁棒性较弱; 此外, 其噪声更易干扰关键特征, 对于简单形状和纹理的保护效果有限. 针对医学图像分割场景, UMed (unlearnable medical image)[52]进一步提出面向医学图像的不可学习方法, 分别在轮廓和纹理两个关键维度进行扰动设计. 在轮廓扰动生成器中, 将普通卷积层替换为中心差分卷积, 以增强对轮廓差异的敏感性, 并利用掩码标注生成轮廓二值图, 从而实现扰动在解剖边界的精确定位, 保证非轮廓区域不受影响. 在纹理扰动生成器中, 首先通过局部二值模式 (local binary pattern, LBP) 提取关键区域的纹理特征图, 随后进行归一化和强度量化; 并引入与纹理强度正相关的自适应扰动边界机制, 使扰动在异质纹理区域显著增强, 而在均质区域趋近于 0. 通过协同注入轮廓感知与纹理感知扰动, UMed 有效提升了医学图像的保护能力. 实验结果表明, 只有当扰动

幅度放宽至特定阈值时, UMed 才能在抵御数据增强与对抗训练方面取得较好效果.

(d) 多任务

基于多类别变换的不可学习方法 (unlearnable approach via multi class-wise transformation, UMT)[53]通过特征空间分析和参数映射机制, 为不同类别的数据样本施加差异化的几何变换策略. 在模型训练过程中, 通过持续监测验证集性能, 动态优化调整变换参数, 确保模型在变换后的数据上学习的决策边界与原始数据分布产生显著偏差, 从而实现数据的不可利用. 该方法生成的不可学习样本在分类和语义分割均有效. UMT 方法适用于 3D 点云数据, 但在图像多任务上有效性未知. Yu 等人[54]进一步研究了现有不可学习样本生成方法在应用于多任务学习场景时的局限性, 并据此提出多任务学习不可学习样本 (multi-task learning unlearnable example, MTL-UE) 框架. 该框架通过一个生成器模型, 将任务特定的标签嵌入与嵌入正则化方案结合, 生成能够同时削弱模型在多个任务上性能的通用扰动. 尽管该方法在多任务下展现出有效性, 但实验结果显示其扰动效果会随任务数量的增加而逐渐减弱. 与图像分类任务不同, 目标检测涉及定位与分类两个子任务, 具有多任务学习的特性. Yi 等人[56]提出了一种基于傅里叶基向量的单模式捷径 (single pattern shortcut, SPS) 方法, 通过在目标边界框内嵌入特定的中频扰动模式, 实现对检测模型的干扰. 该机制使得检测器仅在识别到 SPS 时判定目标存在, 否则将其误判为背景. 实验表明, 该方法显著降低了目标检测模型的性能, 并在多种对抗性防御措施下仍保持较强的鲁棒性.

针对分类任务中涉及身份隐私与任务信息共存的特殊情境, Meng 等人[55]提出脑电图身份不可学习 (identity unlearnable EEG, IUE), 该方法通过最小化任务分类的均方误差, 即保持任务学习轨迹与干净样本一致, 确保任务分类的可学习性; 同时在身份分类任务中引入误差最小化 (EM)[7]思想最小化扰动样本上的交叉熵损失, 从而削弱身份特征的可提取性:

$$\mathcal{L}_{\mathrm{MSE}}\left(H'_C(x+\delta_u), H'_C(x)\right) \tag{6}$$

$$\mathcal{L}_C\left(H'_D(x+\delta_u), y\right) \tag{7}$$

其中, 最小化任务分类 H'_C 在扰动样本和干净样本损失的均方误差, 使得训练轨迹一致, 保持任务分类的可学习性. 对于身份分类任务, 应用误差最小化的思想, 最小化身份分类 H'_D 在扰动样本上的交叉熵损失. 此外, 不可泛化样本 (UGE)[38]也采用类似思路以实现“对授权网络可学习、对非授权网络不可学习”. 具体而言, 该方法利用梯度匹配损失, 使授权网络在扰动样本上的训练轨迹与原始数据一致, 从而保证可学习性; 同时在共享特征空间中计算特征距离损失, 最大化原始样本与不可泛化样本的差异, 阻止非授权网络学习数据. 为防止黑客通过知识蒸馏恢复数据, UGE 设计了不可蒸馏损失, 通过最大化 KL 散度来抑制蒸馏攻击. 实验结果表明, 该方法对知识蒸馏具有较强鲁棒性, 但其性能会随着授权网络数量的增加而有所下降.

(2) 模态多样性方法

模态多样性方法的分类如表 4 所示, 其中, 图数据 (如社交网络、知识图谱、分子结构等) 由节点与边构成, 难以直接应用于图像不可学习方法.

表 4 模态多样性方法分类

模态	方法名称	简介	优点	缺点
图数据	结构误差最小化 (EMinS)[57]	选择梯度绝对值最大的c条边进行修改	具有高效性	未验证对动态图或大规模图的扩展性
3D点云	基于多类别变换的不可学习方法 (UMT)[53]	差异化的几何变换扰动	1) 对分类、分割任务均有效 2) 具有可逆性	无法应用于变换不变的未授权网络
文本	文本不可学 (UT)[58]	选取合适候选词和梯度变换最小的位置进行修改或在末尾添加简单字符或者标点	扩展到文本领域	容易被识别和移除
	ExpShield[49]	干扰模型的token嵌入	具有高效性	极端对抗场景受限
时间序列数据	控制向量噪声 (CVN)[59]	对特定时间段数据添加噪声	根据时间序列数据性质设计	无鲁棒性测试

表 4 模态多样性方法分类 (续)

模态	方法名称	简介	优点	缺点
时间序列数据	脑电图身份不可学习 (IUE)[55]	最小化脑电图任务分类的均方误差和身份分类的交叉熵	有效保护身份隐私的同时保持正常任务分类	小样本下效果有限
音频	HarmonyCloak[51]	利用频率掩蔽技术和窗口策略, 将噪声隐藏在音乐的掩蔽区域内	支持多种音乐格式	未涵盖人声的复杂特性
	音频不可学习 (HiddenSpeaker)[60]	在频域和时域两个维度上优化噪声的不可感知性	具有高效性	无鲁棒性测试
	关键目标扰动 (POP)[50]	重构损失作为扰动优化目标, 实施位置固定的局部扰动策略	1) 计算成本较低 2) 具有可转移性	对抗性攻击的局限性
多模态	多步误差最小化 (MEM)[61]	设计扰动使插入触发器后的文本与扰动图像对比损失最小化	联合优化了图像噪声和文本触发器	文本触发器的生成可能引入语义不连贯的词汇

Liu 等人[57] 提出结构误差最小化 (error-minimizing structural, EMinS), 在邻接矩阵中对所有潜在边计算梯度, 并选取梯度绝对值最大的 c 条边进行修改: 删除具有正梯度的现有边, 或添加具有负梯度的不存在边, 以实现梯度下降并降低分类损失. 同时引入自适应约束, 限制每个顶点可修改的边数, 以及全局修改边数不超过总边数的比例.

3D 点云数据在自动驾驶与虚拟现实等领域的广泛应用, 其中包含的敏感信息面临被未经授权使用的风险. 现有的 2D 图像不可学习方法无法直接迁移至 3D 点云数据, 基于多类别变换的不可学习方法 (UMT)[53]提出基于特征空间解析与参数化映射机制的扰动生成方法, 为不同类别样本注入差异化几何变换. 在训练过程中, 依据验证集性能反馈动态调整参数配置, 并利用几何变换矩阵的可逆性, 使授权用户可通过轻量级参数信息执行逆变换, 恢复原始数据分布以进行训练. 实验表明, 该方法在局部数据增强条件下的扰动效果存在一定衰减.

文本内容并可能预测私人信息, 如敏感话题的情绪. 文本不可学习 (unlearnable text, UT)[58]通过计算候选替代词的稠密向量与原词损失梯度的内积, 选取合适候选词, 并选择梯度变化差异最小的词进行替换. 实验发现替换词多位于文本末尾, 因此提出通过在末尾添加简单字符或标点生成不可学习样本. 然而该模式过于简单, 易被识别和移除, 且部分去偏技术可能消除扰动效果. ExpShield[49]在此基础上引入不可见 token 扰动, 从随机插入逐步过渡到定向攻击低概率 token, 并通过特殊 token (如异常 token、OOV token) 进一步增强干扰.

时间序列数据具有动态和顺序的性质, 难以定位控制噪声的应用区域, 导致扰动效果不均衡. 控制向量噪声 (control vector noise, CVN)[59]设置控制对特定时间段数据添加噪声, 但其鲁棒性不可知. 脑机接口技术通过脑电图 (EEG) 信号实现大脑与外部设备的直接通信, 广泛应用于医疗康复、情绪识别等领域. 然而, EEG 信号不仅包含任务相关信息, 还包含用户身份隐私信息. 为实现任务分类的同时保护用户身份, Meng 等人[55]提出脑电图身份不可学习 (IUE), 提出了两种扰动生成方式: 样本级和用户级. 样本级通过最小化优化任务分类的均方误差和身份分类的交叉熵, 确保身份特征难以提取的同时保证任务分类正常. 用户级扰动为所有用户生成通用扰动, 添加正则化项, 限制扰动幅度. 仅需生成与用户数量相同的扰动, 显著减少计算量.

音频数据具有时序性、频域复杂性及感知敏感性, 需要设计针对特性的不可学习样本生成方法. 针对音乐场景的保护需求, HarmonyCloak[51]以双层优化为框架, 在模型训练阶段最小化生成损失, 并将扰动嵌入音乐的频率隐蔽区域, 在不破坏听觉质量的前提下实现风格仿冒保护. Zhang 等人[60]提出音频不可学习 (HiddenSpeaker), 采用单层误差最小化扰动生成方法, 无需更新模型参数, 大幅降低计算开销. 该方法结合短时傅里叶变换损失与短时客观可懂度损失, 从频域与时域双重优化扰动不可感知性, 并通过加权融合形成感知混合损失函数. 然而其在数据增强场景下的鲁棒性仍需进一步研究. 关键目标扰动 (POP)[50]通过向原始音频中嵌入扰动, 攻击语音合成模型中的重构损失. 对于复杂的分窗生成模型, POP 在固定位置施加局部扰动; 而对轻量级模型, 则采用全局扰动覆盖整个波形.

多模态对比学习依赖海量公开图像-文本数据, 这些数据可能包含敏感的个人隐私信息, 例如社交媒体上的照片和文字描述. 现有的单模态不可学习样本方法在多模态场景下效果有限. Liu 等人[61]提出多步误差最小化 (multi-step error minimization, MEM) 拓展了误差最小化框架, 同时优化了图像噪声和附加的文本触发器. 具体而言, 该方

法采用投影梯度下降优化图像扰动, 并基于梯度替换选择损失下降最快的候选词, 将其作为触发器插入文本, 使得图像与文本扰动联合最小化对比损失. 实验表明, MEM 能有效保护多模态对比学习数据.

2.2 基于启发式的方法不可学习样本生成方法

启发式方法通过显式或隐式引入特定的偏差信息, 协助模型在训练期间寻找“捷径”, 而忽视数据的固有特征. 方法无需依赖替代模型, 能够快速生成扰动. 启发式方法核心思想对比如表 5 所示.

表 5 启发式方法核心思想对比

核心思想	方法名称	简介	优点	缺点
极端像素值	单像素捷径 (OPS))[15]	只修改一个关键像素的值	对对抗训练具有鲁棒性	易受局部数据增强的影响
局部相关性	自回归 (AR) 扰动[16]	使用自回归过程生成扰动	计算成本低	易被滤波器消除
	基于卷积的不可学习数据集 (CUDA)[17]	类特定卷积滤波器模糊图像	计算成本低	模糊参数影响图像质量
	伪恢复投毒 (IRP)[62]	在类别相关卷积后使用线性模型进行不完美恢复	与CUDA相比, 提升了图像质量	计算复杂度高
	基于位置卷积的不可学习音频数据集 (PosCUDA)[63]	对音频特定小块应用类别相关卷积	保持音频质量	依赖于添加扰动音频块的位置
	增强型不可学习 (EUN)样本[64]	双重扰动, 结合全局与局部扰动	提高鲁棒性	模糊参数影响图像质量
线性可分性	线性可分离合成扰动 (LSP)[35]	生成围绕超立方体顶点正态分布的点, 并引入协方差 然后将这些点填充为图像格式	生成速度快	易受灰度化影响
	语义图像深度隐藏 (DH)[28]	嵌入预定义语义图像	语义信息具有鲁棒性	易受JPEG影响
低频信息	影响力引导的水印 (InMark)[65]	将水印嵌入低频空间与具有影响力的像素中	对常见图像压缩具有鲁棒性	计算复杂度高

一般捷径主要干扰整张图片或者明显部分, 单像素捷径 (OPS)[15]对于每个类别的图像, 选择一个固定的像素位置, 并将其修改为相同的目标值. 这个目标值是通过离散搜索找到的, 能够最大程度地偏离原始图像的像素点. 对于每一个类别, 寻找一个最具影响力的像素点, 目标函数公式如下:

$$G_k = \frac{\mathrm{E}[\Delta_k]}{\mathrm{Var}[\Delta_k]} \tag{8}$$

其中, Δ_k 表示颜色变化量, 分子 $\mathrm{E}[\Delta_k]$ 代表颜色变化的平均值, 分母 $\mathrm{Var}[\Delta_k]$ 代表颜色变化的方差. 目标是寻找影响最大且保持稳定的像素点, 需最大化 G_k. 但此方法对于剪切等攻击表现出极强的易被攻击性.

自回归 (Autoregressive, AR) 扰动[16]方法通过使用滑动窗口方法生成扰动, 其中图像的初始部分用高斯噪声填充, 随后的值通过在滑动窗口内使用 AR 过程计算得出, AR 过程公式如下:

$$x_t = \sum_{i=1}^{p} \beta_i x_{t-i} + \varepsilon_t \tag{9}$$

其中, β_i 是自回归系数, 该公式表示当前时间点 t 的值是过去 p 个时间点的值的加权平均值再加上一个误差项 ε_t. 滑动窗口内的应用体现在每个新像素是前 p 个相邻像素的线性组合加上噪声, 有效利用了图像局部区域的空间相关性特征. 自回归扰动不需要访问网络参数, 计算效率显著优于优化方法. 然而实验结果表明, 该方法在面对大半径对抗训练时表现出明显的防御性能下降. 同样利用图像的局部相关性, 基于卷积的不可学习数据集 (CUDA)[17], 通过类特定卷积模糊图像, 使模型学习滤波器与标签的捷径关系. 即每个类别随机生成卷积滤波器, 滤波器的参数由一个私有密钥控制. 对每个类别的图像应用对应的卷积滤波器, 生成模糊化的图像. 对卷积后的图像进行归一化. 通过高斯混合模型的理论分析, 证明了 CUDA 能够有效降低干净数据的分类性能. 若密钥或部分干净数据泄露, 防御效果可能被削弱. 伪恢复投毒 (imperfect restoration poisoning, IRP)[62]基于 CUDA 的滤波框架[17], 首先对图像数据进行初步毒化处理. 然后通过线性回归优化的恢复过程对毒化数据进行改进: 将初步毒化数据分割为局部

块, 针对每个块最小化其中心像素值与线性预测值之间的误差. 最终, 通过将 CUDA 滤波器与学习得到的恢复滤波器进行卷积组合, 生成既能有效破坏模型学习能力, 又能较好保留图像质量的数据. CUDA 通过添加类别模糊噪声使数据集不可学习, 但通常严重降低了数据质量. 相比之下, 基于位置卷积的不可学习音频数据集 (position based convolution for unlearnable audio datasets, PosCUDA)[63]利用 CUDA 方法框架[17], 针对音频数据特有的时序性和多样化特征表示, 仅对局部片段施加噪声, 同时保持其余内容的完整性. 具体而言, 片段位置由类别特定的私钥确定性生成, 随后对选定片段施加类别相关的滤波器. 然而, 该方法对长时音频信号效果有限, 攻击者可能通过提取并利用未受扰动的部分数据来规避保护机制. 为提高生成方法的鲁棒性, 增强型不可学习 (enhanced unlearnable, EUN) 样本[64]结合 OPS[15]与 CUDA[17], 构建双重扰动机制, 迫使模型学习到两种捷径.

模型优化过程中最易捕获的是线性可分离特征. 线性可分离合成扰动 (linear separable synthetic perturbation, LSP)[35]首先通过降维可视化技术和基础分类模型的实验验证, 证明实现有可用性攻击生成的扰动在特征空间中呈现线性可分特性, 并发现扰动是实现学习目标的“捷径”. 线性可分离扰动通过初始生成在高维空间线性可分的数据点, 通过填充操作模拟自然图像的局部相关性, 将所有块拼接并裁剪为原始图像尺寸. 该方法依赖于色彩空间的特征表达, 因此对于灰度化等操作表现出脆弱性. 语义图像深度隐藏 (DH)[28]通过引入预定义的语义图像作为偏差信息, 使模型在训练过程中将注意力集中于隐藏语义而非原始图像的固有特征. 具体而言, DH 首先利用 CLIP 文本提取器对所有文本提示进行 K-means 聚类, 将聚类后的文本提示与边缘地图作为输入, 生成具有隐含语义的语义图像; 随后, 使用可逆神经网络 (INN) 将该语义图像自适应地隐藏到干净图像中.

模型更容易学习到低频信息, Liu 等人[65]提出影响力引导的水印 (influence watermark, InMark), 利用这一特性, 通过离散余弦变换将图像变成在频域空间的表示, 并限制在低频空间生成水印. 利用影响函数计算像素对生成结果的影响力, 寻找关键像素, 指导噪声的添加方向, 确保水印对模型训练过程的破坏性最大化. 在文本嵌入和低秩参数微调 (low-rank adaptation, LoRA) 的数据增强下, 水印仍能阻止模型学习.

2.3 不可学习样本生成方法评价体系

评价不可学习样本的有效性至关重要. 分类任务中的主要评价指标是测试准确率, 即模型在不可学习样本上训练后, 在干净测试集上的测试准确率. 生成任务与分割任务都具有多样的评价指标, 对于图像生成任务, 主要评价指标是弗雷歇初始距离 (Fréchet inception distance, FID) 与身份不匹配分数 (identity mismatch score, IMS). FID 衡量的是生成图像的特征分布与真实图像的特征分布之间的弗雷歇初始距离, 值越小越接近真实图像. IMS 使用 ArcFace 模型提取生成图像和参考干净图像的嵌入特征, 计算余弦相似度. 文本与音频生成根据目标需求设定了不同的评价指标. 不可学习样本生成方法的评价体系总体上主要分为 5 个部分: 有效性、不可感知性、可转移性、鲁棒性和部分数据不可学习.

(1) 有效性: 不可学习样本的目标是防止未授权模型利用数据训练. 在分类任务下, 不可学习样本的有效性表现在用其训练的模型在干净测试集上的准确率显著下降, 接近随机猜测的水平. 通常会与在干净训练样本或已有方法生成的不可学习样本上训练的模型的测试准确率进行对比. 可视化实验能够更加明显地展现数据分布的变化, t-SNE 可视化将高维数据点投影到二维空间中, 对比干净数据与不可学习样本在特征空间中的不同分布直观展示了扰动对干净数据分布的影响. 在生成任务下, 不可学习样本需要阻止模型学习到原始数据的分布或特定内容, 主要评价模型训练后的生成数据与原始数据的相似度. 在分割任务下, 不可学习样本的目的是阻止模型划分出准确的区域, 实验量化其预测分割区域与真实分割区域的重叠程度来验证不可学习样本生成方法的有效性.

(2) 不可感知性: 不可学习样本添加的扰动具有不可感知性, 即不影响数据的正常使用和感官察觉. 对于图像等可直观观察的数据, 实验中将原始样本、不可学习样本以及扰动进行可视化展示. 原始样本与不可学习样本在人眼下应该并无差异. 对于文本, 扰动应不影响连贯性, 使得可以正常被理解含义. 对于音频, 扰动应不影响听觉感受.

(3) 可转移性: 在实际应用过程中, 防御者对攻击者所使用的模型以及数据集存在未知. 需要设计实验验证跨模型架构与数据集的可转移性. 跨模型可转移性的基础实验验证不可学习样本对于不同架构模型的有效性. 在分类任务中, 需要在监督模型的基础上增加对不同无监督模型的有效性测试. 跨数据集可转移性体现在替代模型在

一个数据集上生成的扰动或无替代模型生成的扰动是否可以直接应用在另一个数据集上生成不可学习样本, 实现扰动对不同数据集的保护.

(4) 鲁棒性: 攻击者会采取各种方法来削弱不可学习样本的保护效果, 因此对不可学习样本进行鲁棒性评价是不可缺少的. 实验评价不可学习样本在面对对抗训练、数据预处理、数据增强以及攻击不可学习样本方法后, 其保护效果是否有效.

对抗训练作为提升模型鲁棒性的方法, 不可学习样本的防御效果随着对抗训练的扰动范围增大而逐步减弱. 一般不可学习样本的扰动上限为 8/255, 评价不可学习样本时, 实验会改变对抗训练的强度, 即从 0 开始逐步增强. 分为两种情况, 即对抗扰动小于不可学习样本扰动和对抗扰动大于等于不可学习样本扰动.

数据预处理与数据增强都是从数据层面优化训练深度学习模型的常见策略, 通过对训练数据进行各种变化来提升模型性能. 数据预处理, 例如 JPEG 压缩和灰度化等. JPEG 压缩通过对高频系数进行截断或量化, 可以使图像数据丢失一部分细节信息, 从而实现图像的压缩. 灰度化是指将彩色图像转换为黑白或灰度图像的过程. 在灰度转换过程中, 每个像素点的颜色值被映射到一个介于 0–255 之间的整数值, 表示该像素点的亮度级别. 数据增强, 例如 Mixup、Cutout 等. Mixup 将两个样本的标签进行线性组合, 以此来强制模型学习到更加平滑的决策边界. Cutout 则在图像中随机裁剪出一个区域, 并将其替换为黑色像素, 以此来减少模型对于某些特定区域的关注度. 因此, 实验中会评价不可学习样本在这些策略处理后的鲁棒性. 随着净化不可学习样本方法的研究深入, 在鲁棒性评估时也增加了不可学习对这类方法的防御实验.

(5) 部分数据不可学习: 在现实应用场景下, 当只有部分数据不可学习时, 需要探究不可学习样本生成方法的有效性以及对模型学习剩余干净数据的影响. 实验构建一个包含不可学习样本与原始样本的混合训练集, 测试不可学习样本的不同占比对模型性能的影响.

为了更深入评价不可学习样本, 后续研究开始探索更全面及新的评价方法. 可用性中毒攻击基准 (APBench)[18] 针对分类任务, 构建了一个包含 3 大核心模块的标准化评价框架: 投毒攻击模块、防御模块和评价模块. 该基准整合了 9 种监督学习方法和 2 种无监督学习方法的可用性投毒攻击策略, 同时包含 8 种防御算法和 4 种传统数据增强技术. 在评价方法上, 采用干净测试集准确率作为核心定量指标. 同时, 通过 t-SNE (t-distributed stochastic neighbor embedding) 可视化模型特征表示分布, 并利用 Grad-CAM (gradient-weighted class activation mapping) 和 Shapley Value Map 可视化模型预测过程中关注的图像区域, 从定量和定性两个维度全面评估方法的有效性。

传统的评价方法使用干净测试准确率作为主要评价指标. 虽然能够在一定程度上反映攻击或净化的效果, 但在多任务场景或复杂模型架构下可能存在局限性. 为了更全面地衡量不可学习性, 基于损失景观的不可学习性度量 (unlearnability metrics from loss landscape, UML)[19]提出在多任务场景下不可学习性表现不佳的问题, 从模型优化角度出发, 提出了新的不可学习性评价指标. 通过观察干净样本和不可学习样本在训练过程中的损失景观差异, 发现关键参数的优化路径存在显著分歧, 这一发现证实了损失景观特征与不可学习性之间存在本质关联. 基于此发现, 提出锐度感知可学习性 (sharpness-aware learnability, SAL) 指标, 该指标通过量化参数在损失景观中的平坦程度来评价其可学习性. 具体而言, 较低的 SAL 值表明参数更新对损失函数的影响微乎其微, 即该参数表现出不可学习特性. 基于 SAL 提出不可学习距离 (unlearnable distance, UD), 通过比较干净模型和中毒模型中可学习参数的比例, 量化数据的不可学习性. 实验表明 SAL 与现有不可学习性指标基本一致, 验证了其合理性. 实证显示, 现有不可学习方法在多任务模型中的性能抑制效果有限. 通过观察在不同净化方法以及模型架构下的 UD 值, 发现 JPEG 压缩对不可学习方法净化效果最佳, ViT (vision Transformer) 架构展现出对不可学习方法更强的抵抗能力.

3 不可学习样本净化与检测方法

虽然不可学习样本最初是为了保护个人隐私数据不被深度学习模型在未经授权的情况下利用, 但其本身作为数据投毒攻击的一种方法, 对模型的性能和安全性构成了严重威胁. 通过向数据集中注入不可学习扰动, 可能会对公司的商业利益造成极大损失. 因此, 研究针对不可学习样本的净化技术, 对于模型安全问题来说至关重要. 此外,

净化方法的研究也可以推动不可学习样本的发展, 进一步提升其在数据保护方面的实用性和可靠性.

现有的不可学习样本净化方法可以分为两大类, 分别是模型训练净化和数据预处理净化, 还有一些研究则致力于不可学习样本的检测工作. 模型训练净化指模型直接在不可学习样本上训练, 通过调整训练方案或损失函数使得模型可以从中学到有用的泛化特征, 进而恢复模型在干净测试集上的性能; 数据预处理净化则会在模型训练前对不可学习样本做处理, 以破坏或去除不可学习扰动, 恢复样本的可学习性. 第 3.1 节和 3.2 节将对这两个分类进行详细讨论.

3.1 基于模型训练的不可学习样本净化方法

模型训练净化方法直接让模型在不可学习数据集上训练, 通过修改训练方案使得模型从 UE 中学到泛化特征. 此类方法目前共有 4 种, 其中 3 种均基于对抗训练, 另外 1 种则基于学习率的调整, 其整体框架如图 6 所示.

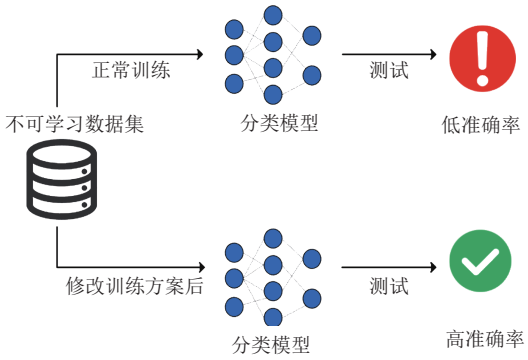

图 6 模型训练净化框架

Tao 等人[26]首次提出使用对抗训练净化不可学习样本, 这也是不可学习样本净化领域的第 1 篇文章. 通过将不可学习样本定义为在一个特定的 ∞-Wasserstein 球中寻找最坏情况下的训练样本, 证明了最小化扰动数据上的对抗训练风险等价于优化原始数据上的自然训练风险上界, 进而得出对抗训练可以作为一种针对不可学习样本的原则性净化这一结论. 为了进一步理解净化的内部机制, 作者在混合高斯分布的简单设置中分析了非鲁棒特征 (non-robust feature) 的作用. 他们发现, 对抗训练可以通过减少模型对这些非鲁棒特征的依赖, 达到抵抗不可学习扰动的效果.

虽然传统的对抗训练可以在一定程度上减轻不可学习样本的影响, 但需要大量的计算资源, 并且会导致模型性能下降. 简单的数据增强作为一种扩充训练数据集、提升模型泛化能力的方法, 并不能净化不可学习样本. UEraser[66]将有效的数据增强策略和损失最大化的对抗性增强方法相结合, 即通常所说的对抗增强 (AA). 该方法使用对抗性增强技术, 突破了当前不可学习攻击和防御所假设的“p 扰动预算”的限制. 它还有助于提升模型的泛化能力, 从而防止准确率下降.

以最大化数据增强样本的误差为核心, Qin 等人[66]设计了一个双层优化目标函数, 如公式 (10) 所示:

$$\arg\min_{\theta}\mathbb{E}_{(x,y)\sim D_{\text{clean}}}\left[\max_{T\sim\mathcal{A}}\mathcal{L}\left(f_{\theta}(T(x)),y\right)\right] \tag{10}$$

其中, 内层的图像转换策略 $\mathcal{T}$ 从所有可能的增强中采样, 以最大限度地增加损失, 而外层使用对抗性策略进行模型训练. 对于从数据集中采样的每个小批量样本, 对每张图像应用 K 种不同的随机增强策略, 并计算所有增强图像对应的损失值. 然后, 选择其 K 种增强变体中产生的最大损失值. 接下来, 使用同一小批量的平均损失值对模型参数进行梯度下降更新. 最后, 完成训练后返回训练好的模型参数. 其中的数据增强策略使用的是 PlasmaTransform 和 TrivialAugment 两套策略, 前者使用基于分形的变换来对图像进行扭曲, 而后者提供一组自然的增强方法. 最后使用 ChannelShuffle 随机交换颜色通道, 进一步增加该方法的攻击性.

由于净化方通常不知道需要应对的不可学习样本的扰动类型 (即范数约束的类型), 而且针对单一类型的扰动进行对抗训练无法应对其他类型的扰动, Zhang 等人[67]提出了一种混合对抗训练策略, 即在每次训练中, 从干扰集

合中均匀采样每种类型的干扰来对数据进行投毒. 该操作在减少计算成本的同时, 可以使模型能够适应现实世界中不可预测的扰动类型, 从而提高模型的泛化能力和鲁棒性. 此外, Zhang 等人[67]还提出类再平衡样本选择 (class-rebalancing sample selection) 策略, 用于应对针对样本标签的攻击. 该文献也是目前唯一一个同时考虑数据损坏和样本损坏的净化方法.

除了借助于对抗训练, Pu 等人[68]提出一种渐进式阶段训练 (progressive staged training, PST) 的训练框架, 实时检测模型是否处于过拟合状态并通过降低学习率减缓学习过程. 他们首先发现模型在训练早期会同时学习不可学习样本中的图像语义特征和扰动特征, 但很快就会对扰动特征过拟合, 尤其是在模型的浅层. 基于这一发现, PST 方法借助于激活簇测量 (activation cluster measurement, *ACM*) 这一指标量化模型是否处于过拟合状态, 其公式如下:

$$ACM(M,D)=\frac{1}{k(k-1)}\sum_{i,j=1}^{k}\frac{L(i,j)}{R(i)\sigma(j)+R(j)\sigma(i)} \tag{11}$$

其中, $L(i,j)$ 代表类别 i 和 j 之间的类间距离, $R(i)$ 和 $R(j)$ 分别代表类别 i 和 j 的半径, $\sigma(i)$ 和 $\sigma(j)$ 则代表类别 i 和 j 的类内距离. 当类间距离变大、类内距离变小时, 意味着模型出现过拟合现象, 对应着 *ACM* 的值超过设定的阈值. 在训练过程中, PST 通过 *ACM* 指标实时监测模型的过拟合倾向. 一旦检测到过拟合迹象, 模型会回退到上一个检查点, 并通过逐步降低浅层的学习率来减缓学习过程, 从而防止模型陷入过拟合. 此外, PST 还结合颜色抖动和灰度化等数据增强技术, 进一步提升模型的学习效果. 这些增强技术使得不可学习扰动更难被模型学习, 从而削弱其对数据保护的能力, 同时也为 PST 方法提供了更复杂的特征空间以避免过拟合. 通过这种逐步调整学习率和数据增强的策略, PST 有效地克服了不可学习样本带来的挑战, 显著提高了模型在不可学习样本上的学习能力.

模型训练净化这类方法执行比较简单, 只需要适当的修改训练方案便可以达到净化不可学习样本的效果. 同时, 这类方法的适用范围广泛, 对多种 UE 均有不错的净化效果. 但由于对抗训练的引入, 使得该类方法需要大量的计算资源, 时间效率不高. 除此之外, 对抗训练还可能会降低模型在干净数据集上的性能.

3.2 基于数据预处理的不可学习样本净化方法

数据预处理净化这类方法在使用搜集到的可疑数据集训练模型之前, 对其进行一些预处理操作, 试图破坏可能存在的不可学习样本, 恢复数据的可学习性, 其框架如后文图 7 所示. 早期的研究通常采用诸如 JPEG 压缩等图像变换操作进行预处理, 随着扩散模型的出现与发展, 一些研究借助扩散模型强大的去除噪声和生成能力, 设计了效果更好且更通用的不可学习样本净化方法.

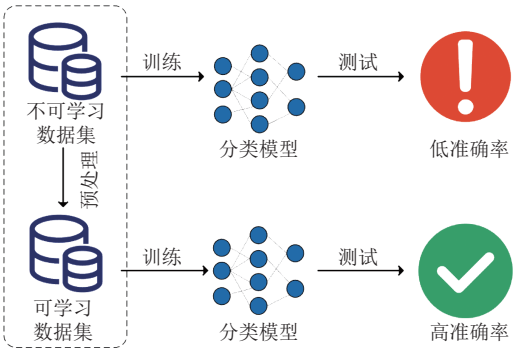

图 7 数据预处理净化框架

3.2.1 基于简单图像变换的净化方法

图像变换操作作为一种常见的预处理数据方法, 一般被用来扩充数据集以提高模型的泛化性. 以往的研究表明, 诸如高斯模糊、随机裁剪、翻转等普通的图像处理方法并不能有效地减轻不可学习样本的影响. 而 JPEG 压缩作为一种常用的图像压缩标准, 通过丢弃图像中的一些细节和信息来实现文件大小的压缩, 可以很好地处理不可学习样本中的捷径, 达到净化的效果. 此外, 混合使用多种非线性变换也可有效地恢复不可学习样本的可学性.

Liu 等人[20]将 JPEG 压缩这一操作引入不可学习样本净化领域中, 其核心思想是压缩图像中因不可学习样本而产生的捷径, 简称图像捷径压缩 (ISS). 通过研究当时的 12 种不可学习样本方法, 并根据不可学习样本是否需要代理模型生成和所使用的代理模型的训练程度, 将其划分为 3 类: 轻度训练的代理模型类 (slightly-trained), 如 DC[39]和 EM[7]; 完全训练的代理模型类 (fully-trained), 如 TC[69]和 TAP[44]; 无需代理模型类 (surrogate-free), 如 LSP[35]和 OPS[15]. 结合频率原则, 即深度神经网络在训练过程中经常从低频率向高频率拟合目标函数, 得出针对轻度训练模型优化的方法能够捕捉到低频模式, 而针对完全训练模型优化的方法则能够捕捉到高频模式这一结论. 对于频率较低但跨颜色通道差异较大的干扰, 采用灰度变换来抑制这些颜色差异. 对于高频干扰, 遵循现有研究中关于消除对抗性干扰的方法, 使用常见的图像压缩操作, 例如 JPEG 压缩和比特位深降低 (bit depth reduction, BDR).

Hapuarachchi 等人[21]提出针对不可学习数据集的非线性变换 (NTAUD), 用于使模型能够从不可学习数据集中学习. 首先对不可学习数据集应用一系列非线性变换, 如膨胀、高斯模糊、阈值二值化、像素操作、旋转和翻转等, 以增强数据的可学习性; 接着, 选择一个预训练的深度学习模型, 并根据需要调整模型架构, 例如添加全连接层和 dropout 层; 然后, 使用经过非线性变换增强的数据集对模型进行训练, 并在验证数据集上评价模型性能. 如果模型出现过拟合或欠拟合的情况, 则相应地调整模型架构、学习率、批量大小和训练周期数等参数, 以优化模型性能; 最终, 当模型在验证集上的准确率超过预设阈值时, 会在测试集上评价模型性能, 从而验证非线性变换框架的有效性.

自监督学习作为一种新兴的人工智能范式, 不需要在训练期间对数据集进行标记, 但这种学习方式也面临数据投毒攻击的威胁. 特别是针对自监督对比学习的对比中毒 (contrastive poisoning) 攻击, 已被证明对多种自监督学习算法具有显著的攻击效果. Guo 等人[70]提出了 Noise-JPEG 方法, 这是一种基于 JPEG 压缩和随机噪声的净化技术, 旨在消除对比中毒攻击对自监督对比学习算法的影响. 在图像预处理阶段, 首先在图像上添加随机噪声, 然后再进行 JPEG 压缩操作, 最后得到处理后的图像.

这类基于简单图像变换的处理方法, 只需要对不可学习数据集做 JPEG 压缩、灰度变换等基本操作, 便可有效地去除不可学习扰动, 且适用范围广泛, 处理成本也很低. 但由于低质量因子的 JPEG 压缩会损坏图像本身的质量, 因此需要在图像质量和不可学习样本净化效果之间寻找一个平衡.

3.2.2 基于图像重生成的净化方法

图像重生成类净化方法利用诸如扩散模型、变分自动编码器 (VAE) 等生成模型对不可学习样本进行重生成, 以去除不可学习扰动. 这类方法不对攻击形式和分类模型做出假设, 因此可以为现有的分类器抵御未知威胁. 然而, 由于早期的生成模型存在模式崩溃、采样质量低等缺点, 并未得到很好的发展. 随着扩散模型的出现, 这类方法迎来了新的发展, 其处理框架如图 8 所示.

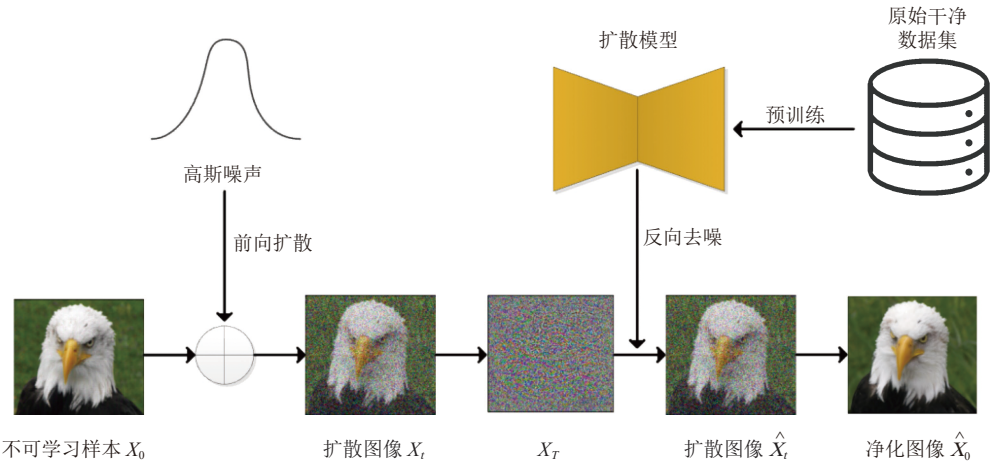

图 8 扩散模型去除不可学习扰动框架

DifffPure[71]首先提出使用扩散模型进行对抗净化: 给出一个对抗样本, 首先在正向扩散过程中使用少量的噪声对其进行扩散, 然后通过反向生成过程恢复干净的图像. 该文献中进行了大量的理论分析, 首先证明了干净的数据分布 $p(x)$ 和受扰动的数据分布 $q(x)$ 在正向扩散过程中会越来越接近, 这意味着对抗性扰动确实会被不断增加的噪声所清理; 接着, 对方法中十分重要的扩散时间步进行分析, 描述了扩散时间步长如何影响净化后得到的图像和原始图像之间的差异, 并得出在设计方法时, 需要仔细权衡 t 的大小, 以确保既能有效去除对抗性扰动, 又能保留图像的语义信息这一结论; 最后, 提出使用伴随方法高效计算梯度, 用于评估较强的自适应攻击. 虽然该方法是为了对抗攻击而设计的, 但也给如何处理不可学习样本带来了新的思路.

在 DiffPure 的基础上, AVATAR[22]将扩散模型应用到不可学习样本领域, 结合扩散模型的正向扩散过程和逆向去噪过程来消除数据保护扰动. 首先, 利用在原始干净数据上预训练的扩散模型对受保护的数据添加适量的高斯噪声, 以减弱数据保护扰动的效果; 然后, 通过扩散模型的逆向过程逐步去除噪声, 恢复数据的原始状态. 通过选择合适的扩散步数, AVATAR 能够在保留数据语义信息的同时, 有效消除保护扰动, 使受保护的数据重新变得可利用. 除此之外, Dolatabadi 等人[22]从理论上证明了用于扩散数据保护扰动所需的噪声量与扰动的大小直接相关. 这一结果表明, 同时实现可用性攻击的两个目标, 即在保证数据可用的前提下实现数据保护是不现实的.

作为 AVATAR 同期的工作, Jiang 等人[23]同样使用扩散模型进行不可学习样本的净化, 并将净化后的样本称作可学习样本 (LE). 在前向扩散阶段, 选择一个相对较小的扩散时间步进行多次迭代比用一个较大的时间步进行一次净化更有效; 为了进一步消除净化强度与语义内容保留之间的权衡, 提出了一种基于扩散模型的新型净化过程, 称为联合条件扩散净化 (joint-conditional diffusion purification, JCDP). JCDP 利用像素距离 (MSE) 和感知距离 (LPIPS) 的双重引导, 实现了对净化图像与原始干净图像之间低层次和高层次语义相似性的联合控制. 与同期使用扩散模型净化不可学习样本的 AVATAR 相比, 该方法不需要使用原始的干净数据进行训练, 只需要分布相同或相近的数据, 增强了实用性; 同时, JCDP 的引入使得净化后的图像更接近原始图像. 而 AVATAR 使用无条件扩散模型净化图像, 随着净化步骤的增加, 生成的图像往往会逐渐偏离原始的干净图像.

无论是 AVATAR 还是 LE, 均需要较多数量的原始干净样本. 而在现实场景中, 模型训练者通常很难得到足够多受保护数据的干净版本, 这限制了该类方法的实际应用. Wang 等人[72]提出稀疏扩散净化 ECLIPSE (expunging clean-label indiscriminate poisons via sparse diffusion purification), 解决了缺少干净数据预训练扩散模型的问题. 假设净化方案仅能获得 5% 的干净数据, 通过直接复制这部分干净数据集, 从而保持增强的数据集分布与原始干净数据集完全一致, 同时也增加了数据量. 将扩散模型在增强后的数据集上进行训练, 并使用其完成对中毒图像的初步净化. 由于低频投毒和鲁棒高频投毒需要更多的高斯噪声同化, 他们设计了一个轻量级的损坏补偿模块来消除这些残留的扰动, 同时确保图像特征不被过度损害. 该模块首先使用概率灰度变换去除残留的低频毒素, 然后再使用轻量级高斯噪声来消除鲁棒的高频毒素.

除了扩散模型, VAE 同样也可用来净化不可学习样本. Yu 等人[73]提出了一种基于速率受限的变分自编码器 (rate-constrained variational autoencoders) 的预训练净化方法. 他们测试了不同 KL 散度目标值对图像重建质量和去除扰动效果的影响, 发现速率受限的 VAE 能够有效地去除 UE 中的扰动, 并通过理论分析揭示了其背后的原理. 除此之外, 还发现了较大的 KL 散度目标值无法有效去除扰动, 而较小的 KL 散度目标值可能会显著降低重建图像的质量. 基于上述发现, 他们提出了一个包含两阶段的净化框架: 第 1 阶段使用较小的 KL 散度目标值, 用不可学习的数据集 P_0 训练 VAE. 该阶段的主要目标是消除数据上大多数的扰动. 在推理过程中, 从 x_0 中减去预测的扰动 $\hat{p}$, 从而获得修改后的图像 p_1; 第 2 阶段则使用较大的 KL 散度目标值进行训练, 以生成高质量的净化数据. 在第 1 次推理中, 从 x_1 中减去预测的扰动 $\hat{p}_1$, 并将其保存为 x_2. 由于扰动是以无监督的方式学习的, 因此很难实现完全重建. 因此, 进行第 2 次推理, 将输出 $\hat{x}_2$ 作为最终结果. 此外, 考虑到大多数扰动都是样本级 (sample-wise) 的, 该工作还引入可学习的类别嵌入 (class-wise embedding), 以更好地预测扰动, 进一步提升净化效果.

图像重生成类净化方法利用了扩散模型强大的去噪和生成能力, 去除不可学习样本上的保护扰动. 此类方法拥有净化效果好, 适用范围广等优点, 且不会影响模型在干净数据集上的性能. 但因为预训练扩散模型需要一定数量的原始干净数据, 限制了其在现实场景中的应用和性能.

3.2.3 其他净化方法

Sandoval-Segura 等人[74]对不可学习样本进行了研究, 发现深度神经网络可以从这类数据中学习到有用特征, 且扰动不一定是线性可分的. 在此基础上, 设计了正交投影攻击方法, 针对类级别、线性可分的扰动. 通过深度特征重加权 (deep feature reweighting, DFR) 方法, 发现受保护数据集训练的网络经 DFR 后, 测试准确率可提升 20%–30%, 说明模型可以从不可学习样本中学到泛化特征; 然后, 用被保护的图像减去干净图像得到对应的净化扰动, 将其用来训练线性逻辑回归模型, 高的训练准确率意味着扰动是线性可分的. 实验结果表明, 在 AR 上的训练准确率只有 39.58%, 说明 AR 的扰动并不是线性可分的, 线性可分并不是不可学习样本的必要条件, 用简单的扰动来定义不可学习样本更为准确; 最后, 提出正交投影攻击 (orthogonal projection attack) 方法, 使用不可学习数据集训练一个线性逻辑回归模型, 优化模型以识别数据中最具有预测性的线性特征, 这些特征通常与类别线性可分的扰动相对应. 通过 QR 分解提取线性模型的权重矩阵 W, 得到正交矩阵 Q. 然后, 将数据集中的每个样本投影到与 Q 正交的空间中, 经过正交投影处理后的数据不再包含原始扰动特征, 从而恢复了数据的可学习性. 最后, 使用这些恢复后的数据重新训练深度神经网络. 实验结果表明, 该方法在类级别扰动上的攻击效果超过了对抗训练, 且计算成本更低.

Li 等人[75]针对基于卷积的不可学习样本, 设计了双线性插值样本破坏方法 COIN. 考虑到卷积不可学习样本的工作原理是误导模型建立类别乘性噪声与真实标签之间的映射关系, 认为可以通过增加类内不一致性并增强类间一致性, 来破坏类别乘性噪声的分布, 进而攻破卷积 UE 的保护. 首先在低维空间设计了净化方法, 通过使用一个随机矩阵 Ar 左乘噪声矩阵, 以此来破坏噪声矩阵的结构, 且通过实验证明了其有效性; 然后, 将低维的方法扩展至真实图像域, 即把乘以 Ar 的操作换为双线性插值, 计算出目标像素点的新像素值, 该过程可看作是对图像中每个像素进行随机加权求和, 从而生成一个新的图像.

同样, Kim 等人[76]提出了一种通过将随机锐化核 (RSK) 和频率滤波 (frequency filtering, FF) 相结合的方法, 使模型从 CUDA[17]中学习到泛化特征. RSK 通过从正态分布中随机采样锐化核的中心值及其邻近值, 打破图像类别级模糊与标签之间的固定关系; 而 FF 则利用离散余弦变换 (DCT) 去除图像中的高频成分, 从而去除由 CUDA 引入的噪声. 实验结果表明, 简单的图像变换可以有效打破基于卷积的不可学习样本保护, 为数据隐私保护技术提出了新的挑战.

3.3 不可学习样本检测方法

相较于不可学习样本净化工作, 检测方面的研究开始的更晚, 且目前正处于初级阶段. 它们通常借助不可学习样本的特性进行判断, 比如线性可分性、模型对于 UE 的快速适应性等, 其框架如图 9 所示. 在对收集到的可疑数据集进行预处理和训练之前, 检测其是否受到不可学习样本的保护具有实际意义, 可以减少不必要的处理, 节省计算资源, 进一步保证模型的安全.

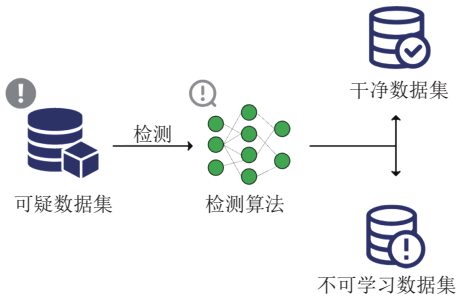

图 9 不可学习样本检测

Dolatabadi 等人[24]提出了一种基于拓扑数据分析 (TDA) 和持久同调 (persistent homology, PH) 的方法来检测和缓解 DNN 中的捷径学习问题, 不可学习样本作为捷径学习的一种, 自然可以用这种方法进行检测. 捷径学习

(shortcut) 会在神经网络的计算图中留下可追踪的路径, 利用持久同调在揭示神经元间连接分支方面的强大表示能力, 可以在包含数千个神经元的高维空间中找到捷径. 首先, 利用神经元的激活向量之间的相关性作为距离度量, 构建一个 Vietoris-Rips (VR) 过滤器. 通过 VR 过滤器, 能够观察到不同粒度下拓扑特征的演变, 并通过持久图 (persistence diagram, PD) 记录这些特征的生命周期. 持久图中的点表示拓扑结构的出生和死亡时间, 而生命周期长的结构更可能是真实数据流形的代表. 他们特别关注了计算图中的 1 维拓扑结构 (即环路), 因为这些结构捕捉到一组高度相关的神经元, 意味着经常被同时激活. 通过分析这些环路, 可以识别出 DNN 中可能存在的捷径路径.

Zhu 等人[77]提出了两种检测不可学习样本的方法: 基于线性可分性的简单网络检测 (simple networks detection, SND) 算法和基于偏移噪声的检测 (bias-shifting noise test, BSNT) 方法. 首先, 通过理论分析证明某些不可学习样本是线性可分的, 而实验表明干净数据集的线性可分率不到 50%. 基于上述发现, 他们提出使用一个线性网络在可疑数据集上进行训练, 如果训练准确率高于设定的阈值 (0.7), 即认为该数据集为中毒数据集, 反之则为干净数据集. 在数据增强的场景下, 由于中毒数据集的线性可分性很容易被破坏, 所以使用一个两层网络进行训练. 除了线性可分性, Zhu 等人[77]还发现不可学习扰动对偏移噪声具有高鲁棒性. 向中毒数据集加入偏移噪声, 模型在验证集上的准确率下降得很少; 而在干净数据集上加入偏移噪声, 会导致准确率大幅度降低, 说明其中的原始特征易被偏移噪声所破坏. 同样, 向可疑数据集的训练集中加入偏移噪声, 并在验证集上测试模型的性能, 如果准确率高于 0.7, 即为中毒数据集. 此外, 他们认为可以通过破坏不可学习样本的可检测性来对抗不可学习样本. 在一个简单的两层网络上进行对抗训练, 然后使用 PGD 攻击生成对抗噪声, 将生成的噪声加入数据集中, 并结合强数据增强训练最终的目标模型. 与传统的对抗训练不同, 该方法只在简单的两层神经网络上进行对抗训练, 大大降低了训练时间.

之前的两种检测方法只能判断模型是否在 UE 上训练, 或者数据集是否被 UE 污染, 不能将一个混合数据集中的干净样本和不可学习样本分离开. Yu 等人[25]提出了一种名为迭代过滤 (IF) 的方法. 该方法的核心在于利用模型在训练过程中对不可学习样本的快速适应特性: 当在一个包含不可学习样本和干净样本的混合数据集上训练分类器时, 模型会优先适应不可学习样本, 导致其在不可学习样本上的准确率比在干净样本上更高. 基于上述发现, IF 方法通过引入额外的类别和迭代细化过程, 逐步分离干净样本和被污染样本. 具体而言, 算法在每次迭代中将数据集分为训练集和验证集, 在训练集上使用早停法训练一个具有扩展类别的分类器, 并根据模型的预测结果更新样本标签. 对于错误预测的样本, 在标签上加上类别总数 C. 由于一些 UE 可能会被错误的分类为干净样本, IF 在几轮迭代之后对干净样本进行一次检索, 以剔除其中的不可学习样本. 随着迭代的进行, 干净样本的标签逐渐被更新为额外类别, 即聚集在 $[C, 2C–1]$ 区间内, 而不可学习样本的标签聚集在 $[0, C–1]$ 区间内, 从而实现两类样本的有效区分.

Li 等人[75]提出了边缘像素检测器 (edge pixel-based detector, EPD), 专门用于基于卷积的不可学习样本的检测. 他们发现基于卷积的 UE 通常会导致图像边缘的像素值较低, 在视觉上体现为出现黑边框. 基于这一发现, EPD 方法首先计算一幅图像 RGB 三通道上的边缘像素值总和, 公式如下:

$$v_d(c) = \left[\sum_{w=1}^{W} e(c,1,w), \sum_{w=1}^{W} e(c,H,w), \sum_{h=1}^{H} e(c,h,1), \sum_{h=1}^{H} e(c,h,w)\right] \tag{12}$$

$$V_d = [v_d(R), v_d(G), v_d(B)] \tag{13}$$

其中, $e(c,h,w)$ 表示在通道 c、行 h 和列 w 处的像素值. 最终的特征向量 V_d 是通过在 3 个通道上拼接 $v_d(c)$ 计算得到的. 将 V_d 作为特征向量输入一个经过预训练的支持向量机 (SVM) 进行二分类, 计算每个输入的决策分数, 如果分数大于设定的阈值, 则认为该图像是经过卷积不可学习样本处理过的, 公式如下:

$$S = w \cdot V_d + b \tag{14}$$

$$\hat{y} = \begin{cases} 1, & \text{if } S(x_d) \geqslant \theta \\ 0, & \text{if } S(x_d) < \theta \end{cases} \tag{15}$$

其中, w 是 SVM 的权重, b 是偏置项. 1 意味着是卷积 UE, 0 则不是.

3.4 现有的不可学习样本净化方法与检测方法的比较

本节对现有的 20 种不可学习样本净化、检测方法进行汇总分析, 分别从普适性、净化性能 (即模型在干净样本上的准确率)、外部依赖性以及算法效率这 4 个方面进行评估, 结果如表 6 所示. 其中的数据依赖分为强、中、弱这 3 个等级, 强即为需要大量的原始干净数据, 中即为需要知道不可学习扰动所使用的范数约束, 弱即为只需要很少的额外数据且不需要知道不可学习样本的类型.

表 6 不可学习样本净化方法

类型	净化方法	净化性能	数据依赖	算法效率	适用范围	方法简述
模型训练净化	对抗训练 (AT)[26]	中	中	低	有约束UE	直接用UE进行对抗训练
	对抗增强 (AA)[66]	高	弱	低	所有UE	结合对抗训练和数据增强
	混合对抗训练 (ERAT)[67]	中	弱	低	有约束UE	从干扰集合中均匀采样每种类型的干扰
	渐进式阶段训练 (PST)[68]	高	弱	低	所有UE	检测模型是否处于过拟合状态, 及时调整学习率
数据预处理净化	图像捷径压缩 (ISS)[20]	高	中	高	所有UE	根据UE类型采用对应的压缩
	针对不可学习数据集的非线性变换 (NTAUD)[21]	中	弱	中	所有UE	对数据集采用各种非线性变换
	Noise-JPEG[70]	高	弱	低	无监督UE	先加随机噪声再进行JPEG压缩
	DiffPure[71]	高	强	低	所有UE	在与原始干净数据分布相近的数据上预训练扩散模型, 使用这个模型对UE先加噪再去噪
	AVATAR[22]	高	强	低		
	可学习样本 (LE)[23]	高	强	低		
	ECLIPSE[72]	低	弱	低		
	D-VAE[73]	高	弱	中	有监督UE	利用速率受限VAE解开UE扰动
	正交投影 (OP)[74]	中	弱	中	类级别UE	在UE上训练逻辑回归模型, 将样本投影到其权重矩阵的正交空间
	COIN[75]	低	弱	高	卷积UE	对图像进行双线性插值操作
	随机锐化核 (RSK)[76]	低	弱	高	卷积UE	先对图像做锐化, 再进行频域滤波
检测	拓扑数据分析 (TDA)[24]	—	弱	高	所有UE	基于模型的持久同调分析
	简单网络检测 (SND)[77]			高		基于UE数据集的线性可分性
	偏移噪声检测 (BSNT)[77]			高		基于UE对偏移噪声的高鲁棒性
	迭代过滤 (IF)[25]			中	有监督UE	基于模型对UE的快速适应性
	边缘像素检测器 (EPD)[75]			高	卷积UE	基于卷积UE边缘像素值偏低

从净化性能来看, 除了基于卷积的不可学习样本的净化方法之外, 其他现有方法在其适用范围内均有不错的效果. 无论是属于模型训练净化的对抗训练和对抗增强方法, 还是属于数据预处理净化的图像捷径压缩和扩散模型净化类方法, 都可以将模型在测试集上的性能恢复到可接受的状态. 而目前的两种净化卷积 UE 方法, 在 CIFAR-10 这一数据集上的准确率均不到 80%, 表明其净化效果欠佳.

在数据依赖方面, 对抗训练和图像捷径压缩需要根据不可学习样本的类型选择相应的约束范数进行训练或选择图像压缩方法进行处理; 而诸如 AVATAR[22]和 LE[23]这类使用扩散模型去除不可学习扰动的方法, 通常需要大量的干净样本预训练扩散模型, 这在现实场景下是难以实现的. 虽然其净化性能出色, 但实际应用的灵活性会受到数据依赖的限制.

从算法效率来看, 模型训练净化这类方法的效率普遍较低, 因为它们大多基于对抗训练, 而对抗训练本身需要大量的计算资源; 数据预处理净化这类方法中的图像变换类分支的效率通常较高, 只要对图像本身做简单的处理, 而由于扩散模型的引入, 图像重生成类分支的效率较低. 算法效率是实际应用中的一个重要考量因素, 高效的净化方法更适合资源有限或需要快速处理的场景.

在适用范围方面, 大部分净化方法都可以处理大多数的不可学习样本, 少部分方法专攻特定的 UE. 比如正交

投影这一方法只适用于类级别、线性可分的 UE; Noise-JPEG 方法被用于处理无监督学习场景下的 UE; COIN 和随机锐化核这两种方法则专门用于处理基于卷积的 UE.

本文还对不可学习样本生成与净化方法的类别、年份和网址进行了整理汇总, 详见表 7.

表 7 不可学习样本生成与净化方法汇总表

类别	方法	年份	网址
生成	针对性对抗投毒 (TAP)	2021	https://github.com/lhfowl/adversarial_poisons
	误差最小化 (EM)	2021	https://github.com/HanxunH/Unlearnable-Examples
	灰度化增强 (GrayAugs)	2021	https://github.com/liuzrcc/ULE-GrayAug
	鲁棒误差最小化 (REM)	2022	https://github.com/fshp971/robust-unlearnable-examples
	可逆变换锁 (LLock)	2022	https://github.com/jinghuichen/Learnability-Lock
	线性可分离合成扰动 (LSP)	2022	https://github.com/dayu11/Availability-Attacks-Create-Shortcuts
	自回归扰动 (AR)	2022	https://github.com/psandovalsegura/autoregressive-poisoning
	文本不可学 (UT)	2022	https://github.com/xinzhel/unlearnable_texts
	簇级特征扰动 (UC)	2022	https://github.com/lhfowl/adversarial_poisons
	自集成保护 (SEP)	2022	https://github.com/Sizhe-Chen/SEP
	单像素捷径 (OPS)	2022	https://github.com/cychomatica/One-Pixel-Shotcut
	可转移不可学习样本 (TUE)	2022	https://github.com/renjie3/TUE
	基于卷积的不可学习数据集 (CUDA)	2023	https://github.com/vinusankars/Convolution-based-Unlearnability
	特征混淆 (EntF)	2023	https://github.com/WenRuiUSTC/EntF
	脑电图不可学习 (IUE)	2023	https://github.com/lbinmeng/EEG_Privacy_Protection
	可证明不可学习样本 (PUE)	2024	https://github.com/NeuralSec/certified-data-learnability
	博弈式不可学习样本 (GUE)	2024	https://github.com/hong-xian/gue
	稳定误差最小化 (SEM)	2024	https://github.com/liuyixin-louis/Stable-Unlearnable-Example
	样本级平均随机性 (ASR)	2024	https://github.com/EuterpeK/Rethinking-Data-Availability-Attacks
	元学习鲁棒扰动 (MetaCloak)	2024	https://github.com/liuyixin-louis/MetaCloak
	多模态概念对齐 (MCA)	2024	https://github.com/xiye7lai/UnlearnableConcept
	分割不可学习 (UnSeg)	2024	https://github.com/sunye23/UnSeg
	基于多类别变换的不可学习方法 (UMT)	2024	https://github.com/CGCL-codes/UnlearnablePC
	增强不可学习示例/增强对抗性毒化 (AUE/AAP)	2024	https://github.com/EhanW/AUE-AAP
	多步误差最小化 (MEM)	2024	https://github.com/thinwayliu/Multimodal-Unlearnable-Examples
	关键目标扰动 (POP)	2024	https://github.com/wxzyd123/Pivotal_Objective_Perturbation
	伪恢复投毒 (IRP)	2024	https://github.com/lyumingzhi/IRP
	扩展不可学习样本 (SUE)	2025	https://github.com/hrlblab/UE_HPC
	多任务学习不可学习样本 (MTL-UE)	2025	https://github.com/yuyi-sd/MTL-UE
评价	可用性中毒攻击基准 (APBench)	2024	https://github.com/lafeat/apbench
	基于损失景观的不可学习性度量 (UML)	2025	https://github.com/MLsecurityLab/HowFarAreFromTrueUnlearnability.git
净化	对抗训练 (AT)	2021	https://github.com/TLMichael/Delusive-Adversary
	DiffPure	2022	https://github.com/NVlabs/DiffPure
	可学习样本 (LE)	2023	https://github.com/jiangw-0/LE_JCDP
	图像捷径压缩 (ISS)	2023	https://github.com/liuzrcc/ImageShortcutSqueezing
	正交投影 (OP)	2023	https://github.com/psandovalsegura/learn-from-unlearnable
	对抗增强 (AA)	2024	https://github.com/lafeat/ueraser
	混合对抗训练 (ERAT)	2024	https://github.com/sduzpf/ERAT
	AVATAR	2024	https://github.com/hmdolatabadi/AVATAR
	ECLIPSE	2024	https://github.com/CGCL-codes/ECLIPSE

表 7　不可学习样本生成与净化方法汇总表 (续)

类别	方法	年份	网址
净化	D-VAE	2024	https://github.com/yuyi-sd/D-VAE
	COIN	2025	https://github.com/wxldragon/COIN
	随机锐化核 (RSK)	2024	https://github.com/aseriesof-tubes/RSK
检测	简单网络检测 (SND)	2024	https://github.com/hala64/udp
	偏移噪声检测 (BSNT)	2024	
	边缘像素检测器 (EPD)	2025	https://github.com/wxldragon/COIN

4　不可学习样本生成与净化方法的未来发展方向

4.1　不可学习样本生成方法未来发展方向

尽管不可学习样本研究取得了显著进展. 但其在鲁棒性、通用性、效率以及应对部分数据不可学习的挑战等方面仍存在显著提升空间. 未来的发展方向可从以下 4 个方面展开.

(1) 提升在对抗训练与数据增强下的鲁棒性

现有方法在对抗训练与数据增强的场景下仍表现出明显脆弱性. 其根本原因在于, 这两类防御机制都会削弱扰动对模型表征学习过程的持续误导能力. 具体而言, 对抗训练通过在最坏扰动邻域内优化模型参数, 促使模型学习更加稳定和鲁棒的判别特征, 从而倾向于抑制、修正甚至忽略不可学习扰动所注入的误导信号; 而数据增强则通过对原始数据进行各种形式的变换或扩展改变样本分布与局部统计特性, 破坏扰动与原始图像之间精细的对应关系, 进而降低扰动的保真性与攻击性能. 当攻击者同时采用对抗训练与数据增强时, 这种双重鲁棒化机制会进一步压缩扰动的有效作用空间, 使保护数据的不可学习能力显著下降, 尤其在对抗训练扰动幅度界限较大时尤为明显.

未来研究可从以下 3 个方向改进. 首先, 从特征空间操纵出发, 而非仅依赖输入空间中基于损失最小化的扰动设计, 使模型在防御训练中学习到的中间表征本身即与真实类别失配, 或被引导至与目标类别无关的错误特征子空间, 从而提升扰动在鲁棒训练条件下的持续误导能力. 其次, 可以探索高频噪声或结构化噪声替代传统像素级 Lp 范数约束下的噪声. 例如, 将扰动能量集中于小波高频子带, 并结合可微 JPEG 压缩训练以提高其在压缩与增强过程中的保持性; 或采用条纹、网格等具有空间结构先验的噪声模式, 以替代独立分布的像素级扰动, 从而增强其对鲁棒训练的抵抗能力. 最后, 还可进一步研究面向优化过程的梯度干预机制. 具体而言, 通过记录代理模型在干净数据上的历史梯度轨迹, 并构造扰动使其诱导的梯度方向与历史平均梯度方向形成显著偏离, 甚至呈大角度冲突, 从而在参数更新阶段持续注入偏离真实优化目标的误导信号, 从优化动力学层面削弱对抗训练的有效性.

(2) 跨任务与跨模态的通用方法

分类任务的不可学习方法主要通过操纵标签预测破坏模型学习; 生成任务的不可学习方法则强调干扰分布建模能力, 以阻止高质量或特定风格内容的生成. 现有方法大多针对单一任务或模态, 缺乏通用性. 未来应探索跨任务不可学习样本, 例如通过多任务联合优化, 针对不同任务分支设计任务专属的扰动适配, 或者借助大规模预训练模型的通用特征表示来设计扰动. 与此同时, 多模态不可学习方法也亟需发展, 其关键在于设计模态自适应或模态无关的噪声捷径: 通过识别不同模态的关键语义表征并建立与扰动的稳定耦合关系, 使扰动在不同模态中均能成为主要学习捷径, 从而实现更强的普适防护能力.

(3) 平衡数据保护效果与计算成本

早期方法通常依赖替代模型指导迭代优化, 计算开销较大; 后续提出的模型无关扰动设计, 在一定程度上缓解了这一问题. 同时, 将扰动从全局缩小至关键局部, 也有效降低了成本. 未来的研究可进一步探索: 1) 通用扰动生成器, 实现一次生成、多数据集迁移使用; 2) 元学习结合偏差信息, 快速生成具有可转移性的扰动; 3) 任务相关性约束, 在确保保护效果的同时控制计算资源消耗.

(4) 解决部分数据不可学习的局限

当数据集中只有部分样本被保护时, 不可学习样本生成方法的整体有效性会因干净样本的存在而显著下降.

未来研究应聚焦于“干净-投毒”混合数据场景下仍然有效的不可学习样本生成机制, 以确保在真实应用中依然具备稳定的防护能力. 具体而言, 可以尝试选择性捷径强化, 即仅对模型不确定性高或代表性强的样本施加扰动, 而非均匀投毒, 从而在相同污染比例下获得更强的整体不可学习效果.

4.2 不可学习样本净化方法未来发展方向

通过对现有不可学习样本净化方法的综合分析, 可以看到当前的研究已经取得了显著进展, 但在实际应用中仍面临诸多挑战. 模型训练净化的方法不需要额外的操作, 直接让模型在不可学习样本上进行训练, 具有适用范围广、处理效果好等优点, 但存在需要大量计算资源且会降低模型在干净数据上的性能等问题; 数据预处理净化方法在模型训练前对不可学习样本进行预处理, 通常需要根据实际情况选择相应的处理方法; 检测方法则通常基于不可学习样本的特征, 比如线性可分性、模型对于 UE 的快速适应性等. 展望未来, 不可学习样本净化方法的研究可以朝着以下 5 个方向发展.

(1) 提高净化方法的通用性

未来亟需发展能够应对多种不可学习样本攻击的通用净化策略, 减少对特定攻击类型的依赖. 其核心在于从不同攻击形式中提炼稳定的共性表征, 构建具有更强泛化能力的净化机制, 这是提升模型整体鲁棒性与安全性的关键方向之一. 现有净化方法大多针对单一攻击范式设计, 例如对抗训练仅对基于梯度优化的扰动有效, 正交投影仅能处理类级别线性可分扰动, 在面对混合攻击或未知攻击时性能会急剧下降. 未来可构建基于扰动特征元学习的通用净化引擎, 通过在典型攻击的大规模元数据集上进行预训练, 学习不同类型不可学习扰动的共性特征表示, 而非针对单一攻击设计特定规则.

(2) 优化算法效率

通过改进算法架构、采用更高效的优化策略或引入轻量化技术, 降低计算复杂度, 提高净化方法的运行效率. 这将使净化方法更适合实时系统和资源受限环境, 同时也能降低净化成本, 提升实际应用的可行性. 现有基于扩散模型的净化方法处理单张图像的时间仍然较长, 基于对抗训练的方法训练周期长达数天, 无法满足大规模数据集处理和实时推理的需求. 未来可探索基于知识蒸馏的轻量化净化网络架构, 将高精度复杂净化模型的知识迁移到轻量级网络中.

(3) 降低数据依赖性

减少对大量干净数据的需求, 开发能够在有限数据条件下实现有效净化的方法. 通过降低数据依赖性, 可以显著提高净化方法的实用性和可扩展性. 现有基于生成模型的净化方法需要数千甚至数万张干净样本预训练模型, 而在医疗、金融等敏感领域, 获取大量标注干净的数据几乎不可能, 未来可发展基于自监督对比学习的少样本净化方法, 利用不可学习样本本身的结构信息和极少量干净样本的先验知识实现有效净化.

(4) 深化对卷积不可学习样本的净化研究

现有的针对卷积不可学习样本的净化方法效果有限, 难以有效应对此类攻击. 未来需要深入研究卷积不可学习样本的特性, 设计更具针对性和有效性的净化策略, 以提升对卷积不可学习样本的净化能力. 可设计基于频域-空域联合的卷积不可学习样本专用净化框架, 针对卷积扰动的频域全局模糊和空域纹理畸变特性进行双域协同处理.

(5) 拓展净化方法的应用范围

目前的不可学习样本净化方法主要集中在图像领域, 而音频、视频、文本等其他领域也面临着类似的不可学习样本威胁. 未来的研究应将净化方法拓展到这些领域, 开发适用于音频、视频、文本等多模态数据的净化技术, 以全面应对不同领域的不可学习样本攻击. 可构建模态自适应的跨模态通用净化框架, 设计与模态无关的核心扰动检测引擎, 结合各模态的特性进行专属适配.

References

[1] Liu RX, Chen H, Guo RY, Zhao D, Liang WJ, Li CP. Survey on privacy attacks and defenses in machine learning. Ruan Jian Xue Bao/Journal of Software, 2020, 31(3): 866–892 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5904.htm [doi: 10.13328/j.cnki.jos.005904]

[2] Sina Finance. Dior punished by public security! Cartier and LV also suffered customer data leaks. 2025 (in Chinese). https://finance.sina.com.cn/roll/2025-09-09/doc-infpwvhn0441445.shtml

[3] Birhane A, Prabhu VU. Large image datasets: A pyrrhic win for computer vision? In: Proc. of the 2021 IEEE Winter Conf. on Applications of Computer Vision. Waikoloa: IEEE, 2021. 1536–1546. [doi: 10.1109/WACV48630.2021.00158]

[4] Wang ZB, Wang X, Ma JJ, Qin Z, Ren J, Ren K. Survey on adversarial example attack for computer vision systems. Chinese Journal of Computers, 2023, 46(2): 436–468 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2023.00436]

[5] Gao MN, Chen W, Wu LF, Zhang BL. Survey on backdoor attacks and defenses for deep learning research. Ruan Jian Xue Bao/Journal of Software, 2025, 36(7): 3271–3305 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/7364.htm [doi: 10.13328/j.cnki.jos.007364]

[6] Zhang ZH, Fu Y, Gao TG. Research on federated deep neural network model for data privacy preserving. Acta Automatica Sinica, 2022, 48(5): 1273–1284 (in Chinese with English abstract). [doi: 10.16383/j.aas.c200236]

[7] Huang HX, Ma XJ, Erfani SM, Bailey J, Wang YS. Unlearnable examples: Making personal data unexploitable. In: Proc. of the 9th Int'l Conf. on Learning Representations. OpenReview.net, 2021.

[8] Wang ZR, Wang YF, Wang YS. Fooling adversarial training with inducing noise. arXiv:2111.10130, 2021.

[9] Liu ZR, Zhao ZY, Kolmus A, Berns T, van Laarhoven T, Heskes T, Larson M. Going grayscale: The road to understanding and improving unlearnable examples. arXiv:2111.13244, 2021.

[10] Wen R, Zhao ZY, Liu ZR, Backes M, Wang TH, Zhang Y. Is adversarial training really a silver bullet for mitigating data poisoning? In: Proc. of the 11th Int'l Conf. on Learning Representations. Kigali: OpenReview.net, 2023.

[11] Ren J, Xu H, Wan YX, Ma XJ, Sun LC, Tang JL. Transferable unlearnable examples. In: Proc. of the 11th Int'l Conf. on Learning Representations. Kigali: OpenReview.net, 2023.

[12] Sun WX, Liu YX, Yan ZL, Xu KD, Sun LC. Medical unlearnable examples: Securing medical data from unauthorized training via sparsity-aware local masking. In: Proc. of the 41st Int'l Conf. on Machine Learning. Vienna: PMLR, 2024. 1–11.

[13] Zhao ZY, Duan JH, Hu X, Xu KD, Wang CA, Zhang R, Du ZD, Guo Q, Chen YJ. Unlearnable examples for diffusion models: Protect data from unauthorized exploitation. In: Proc. of the ICLR 2024 Workshop on Reliable and Responsible Foundation Models. Vienna: OpenReview.net, 2024.

[14] Sun Y, Zhang H, Zhang TH, Ma XJ, Jiang YG. UnSeg: One universal unlearnable example generator is enough against all image segmentation. In: Proc. of the 38th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2024. 2513

[15] Wu ST, Chen SZ, Xie CH, Huang XL. One-pixel shortcut: On the learning preference of deep neural networks. In: Proc. of the 11th Int'l Conf. on Learning Representations. Kigali: OpenReview.net, 2023.

[16] Sandoval-Segura P, Singla V, Geiping J, Goldblum M, Goldstein T, Jacobs DW. Autoregressive perturbations for data poisoning. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 1985.

[17] Sadasivan VS, Soltanolkotabi M, Feizi S. CUDA: Convolution-based unlearnable datasets. In: Proc. of the 43rd IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 3862–3871. [doi: 10.1109/CVPR52729.2023.00376]

[18] Qin TR, Gao XT, Zhao JJ, Ye KJ. , Xu CZ. APBench: A unified availability poisoning attack and defenses benchmark. Trans. on Machine Learning Research, 2024.

[19] Ye K, Su LC, Qian CX. How far are we from true unlearnability? In: Proc. of the 13th Int'l Conf. on Learning Representations. Singapore: OpenReview.net, 2025.

[20] Liu ZR, Zhao ZY, Larson M. Image shortcut squeezing: Countering perturbative availability poisons with compression. In: Proc. of the 40th Int'l Conf. on Machine Learning. Honolulu: PMLR, 2023. 22473–22487.

[21] Hapuarachchi T, Lin J, Xiong KQ, Rahouti M, Ost G. Nonlinear transformations against unlearnable datasets. arXiv:2406.02883, 2026.

[22] Dolatabadi HM, Erfani S, Leckie C. The devil's advocate: Shattering the illusion of unexploitable data using diffusion models. In: Proc. of the 3rd IEEE Conf. on Secure and Trustworthy Machine Learning. Toronto: IEEE, 2024. 358–386. [doi: 10.1109/SaTML59370.2024.00025]

[23] Jiang W, Diao YF, Wang H, Sun JX, Wang M, Hong RC. Unlearnable examples give a false sense of security: Piercing through unexploitable data with learnable examples. In: Proc. of the 31st ACM Int'l Conf. on Multimedia. Ottawa: ACM, 2023. 8910–8921. [doi: 10.1145/3581783.3611833]

[24] Dolatabadi HM, Erfani SM, Leckie C. Be persistent: Towards a unified solution for mitigating shortcuts in deep learning. In: Proc. of the 27th European Conf. on Artificial Intelligence. Santiago de Compostela: IOS Press, 2024. 2540–2547. [doi: 10.3233/FAIA240783]

[25] Yu Y, Zheng QC, Yang SY, Yang WH, Liu J, Lu SJ, Tan YP, Lam KY, Kot A. Unlearnable examples detection via iterative filtering. In: Proc. of the 33rd Int'l Conf. on Artificial Neural Networks and Machine Learning. Lugano: Springer, 2024. 241–256. [doi: 10.1007/978-3-

031-72359-9_18]

[26] Tao L, Feng L, Yi JF, Huang SJ, Chen SC. Better safe than sorry: Preventing delusive adversaries with adversarial training. In: Proc. of the 35th Int'l Conf. on Neural Information Processing Systems. Curran Associates Inc., 2021. 1240.

[27] Fu SP, He FX, Liu Y, Shen L, Tao DC. Robust unlearnable examples: Protecting data privacy against adversarial learning. In: Proc. of the 10th Int'l Conf. on Learning Representations. OpenReview.net, 2022. 1–22.

[28] Meng RH, Yi CY, Yu Y, Yang SY, Shen BQ, Kot AC. Semantic deep hiding for robust unlearnable examples. IEEE Trans. on Information Forensics and Security, 2024, 19: 6545–6558. [doi: 10.1109/TIFS.2024.3421273]

[29] Fang B, Li B, Wu S, Ding SH, Yi R, Ma LZ. Re-thinking data availability attacks against deep neural networks. In: Proc. of the 47th IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 12215–12224. [doi: 10.1109/CVPR52733.2024.01161]

[30] Liu YX, Xu KD, Chen X, Sun LC. Stable unlearnable example: Enhancing the robustness of unlearnable examples via stable error-minimizing noise. In: Proc. of the 38th AAAI Conf. on Artificial Intelligence. Vancouver: AAAI Press, 2024. 3783–3791. [doi: 10.1609/aaai.v38i4.28169]

[31] Zhang YS. Deep clustering based construction method of unlearnable examples [MS. Thesis]. Chengdu: University of Electronic Science and Technology of China, 2024 (in Chinese with English abstract). [doi: 10.27005/d.cnki.gdzku.2024.001199]

[32] Wang YH, Zhu YF, Gao XS. Efficient availability attacks against supervised and contrastive learning simultaneously. In: Proc. of the 38th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2024. 2320.

[33] Gong XL, Wang YJ, Chen YJ, Dong HC, Li YM, Sun MY, Li SK, Wang Q. ARMOR: Shielding unlearnable examples against data augmentation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2026, 48(4): 3988–4004. [doi: 10.1109/TPAMI.2026.3652456]

[34] Zhang ZL, Zhang J, Zhang K, Zhou WB, Xu T, Gao DH, Guo ZX, Guo QL, Zhang WM, Yu NH. Segue: Side-information guided generative unlearnable examples for facial privacy protection in real world. In: Proc. of the 50th IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. Hyderabad: IEEE, 2025. 1–5. [doi: 10.1109/ICASSP49660.2025.10889952]

[35] Yu D, Zhang HS, Chen W, Yin J, Liu TY. Availability attacks create shortcuts. In: Proc. of the 28th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining. Washington: ACM, 2022. 2367–2376. [doi: 10.1145/3534678.3539241]

[36] Zhang JM, Ma XJ, Yi Q, Sang JT, Jiang YG, Wang YW, Xu CS. Unlearnable clusters: Towards label-agnostic unlearnable examples. In: Proc. of the 44th IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 3984–3993. [doi: 10.1109/CVPR52729.2023.00388]

[37] Zhu YF, Lyngaas I, Meena MG, Koran MEI, Malin B, Moyer D, Bao SX, Kapadia A, Wang X, Landman B, Huo YK. Scale-up unlearnable examples learning with high-performance computing. Electronic Imaging, 2025, 37: 184-1–184-6. [doi: 10.2352/EI.2025.37.12.HPCI-184]

[38] Ye JW, Wang XC. Ungeneralizable examples. In: Proc. of the 45th IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 11944–11953. [doi: 10.1109/CVPR52733.2024.01135]

[39] Fang B, Li B, Wu S, Zheng TY, Ding SH, Yi R, Ma LZ. Towards generalizable data protection with transferable unlearnable examples. arXiv:2305.11191, 2023.

[40] Chen CC, Zhang JM, Li YY, Han ZX. One for all: A universal generator for concept unlearnability via multi-modal alignment. In: Proc. of the 41st Int'l Conf. on Machine Learning. Vienna: PMLR, 2024. 7700–7711.

[41] Liu S, Wang YH, Gao XS. Game-theoretic unlearnable example generator. In: Proc. of the 38th AAAI Conf. on Artificial Intelligence. Vancouver: AAAI Press, 2024. 21349–21358. [doi: 10.1609/aaai.v38i19.30130]

[42] Wang DR, Xue MH, Li B, Camtepe S, Zhu LM. Provably unlearnable data examples. In: Proc. of the 32nd Annual Network and Distributed System Security Sympo. San Diego: The Internet Society, 2025. 1–18. [doi: 10.14722/ndss.2025.240886]

[43] Peng WQ, Chen JH. Learnability lock: Authorized learnability control through adversarial invertible transformations. In: Proc. of the. 10th Int'l Conf. on Learning Representations. OpenReview.net, 2022. 1–19.

[44] Fowl L, Goldblum M, Chiang PY, Geiping J, Czaja W, Goldstein T. Adversarial examples make strong poisons. In: Proc. of the 35th Int'l Conf. on Neural Information Processing Systems. New York: Curran Associates Inc., 2021. 2321.

[45] Chen SZ, Yuan G, Cheng XW, Gong YF, Qin MH, Wang YZ, Huang XL. Self-ensemble protection: Training checkpoints are good data protectors. In: Proc. of the 11th Int'l Conf. on Learning Representations. Kigali: OpenReview.net, 2023. 1–12.

[46] Liu YX, Fan CR, Dai YT, Chen X, Zhou P, Sun LC. MetaCloak: Preventing unauthorized subject-driven text-to-image diffusion-based synthesis via meta-learning. In: Proc. of the 45th IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 24219–24228. [doi: 10.1109/CVPR52733.2024.02286]

[47] Wu ZH, Cheng YS, Sun TY, Ji XY, Xu WY. MYOPIA: Protecting face privacy from malicious personalized text-to-image synthesis via

unlearnable examples. In: Proc. of the 39th AAAI Conf. on Artificial Intelligence. Philadelphia: AAAI Press, 2025. 905–913. [doi: 10.1609/aaai.v39i1.32075]

[48] Zhou X, Lu Y, Ma RT, Wei YJ, Gui T, Zhang Q, Huang XJ. Making harmful behaviors unlearnable for large language models. In: Findings of the Association for Computational Linguistics: ACL 2024. Bangkok: ACL, 2024. 10258–10273. [doi: 10.18653/v1/2024.findings-acl.611]

[49] Liu RX, Tran T, Wang TH, Hu HS, Wang S, Xiong L. ExpShield: Safeguarding Web text from unauthorized crawling and language modeling exploitation. arXiv:2412.21123v1, 2024.

[50] Zhang ZS, Yang QY, Wang DR, Huang PY, Cao YX, Ye K, Hao J. Mitigating unauthorized speech synthesis for voice protection. In: Proc. of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis. Salt Lake City: ACM, 2024. 13–24. [doi: 10.1145/3689217.3690615]

[51] Meerza SIA, Sun LC, Liu J. HarmonyCloak: Making music unlearnable for generative AI. In: Proc. of the 46th IEEE Symp. on Security and Privacy (SP). San Francisco: IEEE, 2025. 430–448. [doi: 10.1109/SP61157.2025.00085]

[52] Lin X, Yu Y, Xia S, Jiang J, Wang HR, Yu ZT, Liu YZ, Fu Y, Wang S, Tang WZ, Kot A. Safeguarding medical image segmentation datasets against unauthorized training via contour-and texture-aware perturbations. arXiv:2403.14250, 2024.

[53] Wang XL, Li MH, Liu W, Zhang HT, Hu SS, Zhang YC, Zhou ZQ, Jin H. Unlearnable 3D point clouds: Class-wise transformation is all you need. In: Proc. of the 38th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2024. 3154.

[54] Yu Y, Xia S, Yang SY, Kong CQ, Yang WH, Lu SJ, Tan YP, Kot AC. MTL-UE: Learning to learn nothing for multi-task learning. In: Proc. of the 42nd Int'l Conf. on Machine Learning. Vancouver: PMLR, 2025. 73286–73303.

[55] Meng LB, Jiang X, Huang J, Li W, Luo HB, Wu DR. User identity protection in EEG-based brain—computer interfaces. IEEE Trans. on Neural Systems and Rehabilitation Engineering, 2023, 31: 3576–3586. [doi: 10.1109/TNSRE.2023.3310883]

[56] Yi CY, Meng RH, Peng HH, Shen BQ, Kot AC. Towards effective and robust unlearnable examples against object detection. In: Proc. of the 32nd IEEE Int'l Conf. on Image Processing. Anchorage: IEEE, 2025. 91–96. [doi: 10.1109/ICIP55913.2025.11084567]

[57] Liu YX, Fan CR, Zhou P, Sun LC. Unlearnable graph: Protecting graphs from unauthorized exploitation. arXiv:2303.02568, 2023.

[58] Li XZ, Liu M. Make text unlearnable: Exploiting effective patterns to protect personal data. In: Proc. of the 3rd Workshop on Trustworthy Natural Language Processing. Toronto: ACL, 2023. 249–259. [doi: 10.18653/v1/2023.trustnlp-1.22]

[59] Jiang YJ, Ma XJ, Erfani SM, Bailey J. Unlearnable examples for time series. In: Proc. of the 28th Pacific-Asia Conf. on Knowledge Discovery and Data Mining. Taipei: Springer, 2024. 213–225. [doi: 10.1007/978-981-97-2266-2_17]

[60] Zhang ZS, Huang PY. HiddenSpeaker: Generate imperceptible unlearnable audios for speaker verification system. In: Proc. of the 33rd Int'l Joint Conf. on Neural Networks. Yokohama: IEEE, 2024. 1–8. [doi: 10.1109/IJCNN60899.2024.10651155]

[61] Liu XW, Jia XJ, Xun Y, Liang SY, Cao XC. Multimodal unlearnable examples: Protecting data against multimodal contrastive learning. In: Proc. of the 32nd ACM Int'l Conf. on Multimedia. Melbourne: ACM, 2024. 8024–8033. [doi: 10.1145/3664647.3680708]

[62] Huang Y, Styborski J, Lyu MZ, Wang F, Kong A. Leveraging imperfect restoration for data availability attack. In: Proc. of the 18th European Conf. on Computer Vision. Milan: Springer, 2024. 69–86. [doi: 10.1007/978-3-031-73464-9_5]

[63] Gokul V, Dubnov S. PosCUDA: Position based convolution for unlearnable audio datasets. arXiv:2401.02135, 2024.

[64] Chen XT, Xu Y, Zhang SC, Yan JL, Xu WD, He XL. EUN: Enhanced unlearnable examples generation approach for privacy protection. Computer Vision and Image Understanding, 2025, 258: 104388. [doi: 10.1016/j.cviu.2025.104388]

[65] Liu HW, Sun ZC, Mu YD. Countering personalized text-to-image generation with influence watermarks. In: Proc. of the 37th IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 12257–12267. [doi: 10.1109/CVPR52733.2024.01165]

[66] Qin TR, Gao XT, Zhao JJ, Ye KJ, Xu CZ. Learning the unlearnable: Adversarial augmentations suppress unlearnable example attacks. In: Proc. of the 12th Int'l Conf. on Learning Representations. Vienna: OpenReview.net, 2024. 1–19.

[67] Zhang PF, Huang Z, Xu XS, Bai GD. Effective and robust adversarial training against data and label corruptions. IEEE Trans. on Multimedia, 2024, 26: 9477–9488. [doi: 10.1109/TMM.2024.3394677]

[68] Dang PC, Hu X, Xu KD, Duan JH, Huang D, Han HS, Zhang R, Du ZD, Guo Q, Chen YJ. Flew over learning trap: Learn unlearnable samples by progressive staged training. arXiv:2306.02064v1, 2023.

[69] Shen JC, Zhu XL, Ma D. TensorClog: An imperceptible poisoning attack on deep neural network applications. IEEE Access, 2019, 7: 41498–41506. [doi: 10.1109/ACCESS.2019.2905915]

[70] Guo WH, Ma XJ, Ge PY, Chen Y, Yue QL, Zhang YQ. Defense contrastive poisoning: An application of JPEG to self-supervised contrastive learning indiscriminate poisoning attacks. In: Proc. of the 17th IEEE Int'l Conf. on Internet of Things. Copenhagen: IEEE, 2024. 505–510. [doi: 10.1109/iThings-GreenCom-CPSCom-SmartData-Cybermatics62450.2024.00097]

[71] Nie WL, Guo B, Huang YJ, Xiao CW, Vahdat A, Anandkumar A. Diffusion models for adversarial purification. In: Proc. of the 39th Int'l

Conf. on Machine Learning. Baltimore: PMLR, 2022. 16805–16827.

[72] Wang XL, Hu SS, Zhang YC, Zhou ZQ, Zhang LY, Xu P, Wan W, Jin H. ECLIPSE: Expunging clean-label indiscriminate poisons via sparse diffusion purification. In: Proc. of the 29th European Symp. on Research in Computer Security. Bydgoszcz: Springer, 2024. 146–166. [doi: 10.1007/978-3-031-70879-4_8]

[73] Yu Y, Wang YF, Xia S, Yang WH, Lu SJ, Tan YP, Kot A. Purify unlearnable examples via rate-constrained variational autoencoders. In: Proc. of the 41st Int'l Conf. on Machine Learning. Vienna: PMLR, 2024. 57678–57702.

[74] Sandoval-Segura P, Singla V, Geiping J, Goldblum M, Goldstein T. What can we learn from unlearnable datasets? In: Proc. of the 37th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2023. 3294.

[75] Li MH, Wang XL, Yu ZF, Hu SS, Zhou ZQ, Zhang LL, Zhang LY. Detecting and corrupting convolution-based unlearnable examples. In: Proc. of the 39th AAAI Conf. on Artificial Intelligence. Philadelphia: AAAI Press, 2025. 18403–18411. [doi: 10.1609/aaai.v39i17.34025]

[76] Kim D, Sandoval-Segura P. Learning from convolution-based unlearnable datasets. In: Proc. of the 3rd Workshop on New Frontiers in Adversarial Machine Learning. Vancouver: OpenReview.net, 2024. 1–10.

[77] Zhu YF, Yu LJ, Gao XS. Detection and defense of unlearnable examples. In: Proc. of the 38th AAAI Conf. on Artificial Intelligence. Vancouver: AAAI Press, 2024. 17211–17219. [doi: 10.1609/aaai.v38i15.29667]

附中文参考文献

[1] 刘睿瑄, 陈红, 郭若杨, 赵丹, 梁文娟, 李翠平. 机器学习中的隐私攻击与防御. 软件学报, 2020, 31(3): 866–892. http://www.jos.org.cn/1000-9825/5904.htm [doi: 10.13328/j.cnki.jos.005904]

[2] 新浪财经. 迪奥遭公安处罚! 客户数据泄露的还有卡地亚和 LV. 2025. https://finance.sina.com.cn/roll/2025-09-09/doc-infpwvhn0441445.shtml

[4] 王志波, 王雪, 马菁菁, 秦湛, 任炬, 任奎. 面向计算机视觉系统的对抗样本攻击综述. 计算机学报, 2023, 46(2): 436–468. [doi: 10.11897/SP.J.1016.2023.00436]

[5] 高梦楠, 陈伟, 吴礼发, 张伯雷. 面向深度学习的后门攻击及防御研究综述. 软件学报, 2025, 36(7): 3271–3305. http://www.jos.org.cn/1000-9825/7364.htm [doi: 10.13328/j.cnki.jos.007364]

[6] 张泽辉, 富瑶, 高铁杠. 支持数据隐私保护的联邦深度神经网络模型研究. 自动化学报, 2022, 48(5): 1273–1284. [doi: 10.16383/j.aas.c200236]

[31] 张弈晟. 基于深度聚类的不可学习样本构造方法 [硕士学位论文]. 成都: 电子科技大学, 2024. [doi: 10.27005/d.cnki.gdzku.2024.001199]

作者简介

孟若涵, 博士, 讲师, 主要研究领域为不可学习样本, 数据投毒, 数据安全.

卞新玉, 硕士生, 主要研究领域为不可学习样本, 隐私保护.

张丛阳, 本科生, 主要研究领域为不可学习样本.

朱浩天, 本科生, 主要研究领域为不可学习样本, 信息隐藏.

付章杰, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为数据安全, 外包安全, 数字取证, 网络信息安全.