软件学报 ISSN 1000-9825, CODEN RUXUEW
[doi: 10.13328/j.cnki.jos.007673] [CSTR: 32375.14.jos.007673]

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

基于数据重建的净标签数据投毒攻击方案*

张文博[1], 李 雄[1,2], 张文琪[1], 谢 勇[3], 陈 厅[1,2], 禹继国[1], 张小松[1,2]

[1](电子科技大学 计算机科学与工程学院, 四川 成都 611731)

[2](电子科技大学 (深圳) 高等研究院, 广东 深圳 518110)

[3](南京邮电大学 计算机学院, 江苏 南京 210023)

通信作者: 李雄, E-mail: lixiong@uestc.edu.cn; 谢勇, E-mail: yongxie@njupt.edu.cn

摘 要: 人工智能系统依赖大规模、开放式的训练数据采集, 这为数据投毒攻击提供了可乘之机. 攻击者能够通过渗透数据供应链等方式操控训练数据, 破坏深度学习模型的可用性. 相较于基于标签翻转的脏标签数据投毒攻击, 净标签攻击虽具有更强的隐蔽性, 但在实现时面临更大挑战: 一方面, 需要精确操控样本特征而非直接篡改其标签; 另一方面, 还需确保毒化样本在视觉上与原始样本保持高度一致以规避检测. 虽然现有的净标签攻击方案已取得一定成效, 但往往因假设宽松, 难以在实际场景中有效实施, 且隐蔽性和计算效率仍有提升空间. 为此, 提出一种基于数据重建的深度学习净标签数据投毒攻击方案 CPDR, 该方案重构原用于图像分割任务的 Attention U-Net 模型的损失函数, 将其转换为一个包含重建损失、投毒效果损失和模型更新一致性损失的多任务学习重建模型, 通过数据重建生成隐蔽的毒化样本, 并在黑盒假设下实现有效的投毒攻击. 在集中式学习和联邦学习这 2 种场景、2 个数据集和 4 个目标模型上的实验表明, CPDR 能在几乎不影响深度学习主任务表现的情况下降低模型在目标类别上的分类准确率. 在集中式学习下, CPDR 对目标类别准确率的平均降幅高于 VagueGAN 方案 2.33 个百分点, 高于 Adversarial Poison 方案 2.42 个百分点; 在联邦学习下, 其优势进一步扩大, 分别高于上述对比方案 4.44 和 4.10 个百分点. 此外, CPDR 生成的毒化样本与原始样本视觉差异微小, 难以辨别和检测, 实现了隐蔽性和攻击效果的良好平衡.

关键词: 数据投毒; 投毒攻击; 净标签攻击; 人工智能安全

中图法分类号: TP309

中文引用格式: 张文博, 李雄, 张文琪, 谢勇, 陈厅, 禹继国, 张小松. 基于数据重建的净标签数据投毒攻击方案. 软件学报. http://www.jos.org.cn/1000-9825/7673.htm

英文引用格式: Zhang WB, Li X, Zhang WQ, Xie Y, Chen T, Yu JG, Zhang XS. Clean-label Data Poisoning Attack Scheme Based on Data Reconstruction. Ruan Jian Xue Bao/Journal of Software (in Chinese). http://www.jos.org.cn/1000-9825/7673.htm

Clean-label Data Poisoning Attack Scheme Based on Data Reconstruction

ZHANG Wen-Bo[1], LI Xiong[1,2], ZHANG Wen-Qi[1], XIE Yong[3], CHEN Ting[1,2], YU Ji-Guo[1], ZHANG Xiao-Song[1,2]

[1](School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China)

[2](Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Shenzhen 518110, China)

[3](School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

Abstract: Artificial intelligence systems rely on large-scale, open training data collection, which creates opportunities for data poisoning attacks. Attackers can manipulate training data by infiltrating the data supply chain, thus compromising the availability of deep learning models. Compared to dirty-label data poisoning attacks based on label flipping, clean-label attacks are more stealthy but face greater challenges in implementation. On the one hand, precise manipulation of sample features is required rather than direct alteration of labels.

* 基金项目: 国家自然科学基金 (U25A20429, 62332018, 62332004); 中央高校基本科研业务费专项资金 (ZYGX2025TS010); 四川省自然科学基金 (2025ZNSFSC0497); 深圳市科技重大专项 (KJZD20231023092900002)
收稿时间: 2025-07-07; 修改时间: 2025-09-18, 2026-02-26; 采用时间: 2026-03-23; jos 在线出版时间: 2026-07-08

On the other hand, it is necessary to ensure that poisoned samples remain highly visually consistent with the original samples to evade detection. Although existing clean-label attacks have achieved some success, they are often difficult to implement in practical scenarios due to loose assumptions, and there remains room for improvement in terms of stealthiness and computational efficiency. To address these issues, this study proposes a clean-label data poisoning attack scheme for deep learning based on data reconstruction, termed CPDR. In this scheme, the loss function of the Attention U-Net model, originally designed for image segmentation, is reconstructed and transformed into a multi-task learning reconstruction model that incorporates reconstruction loss, attack effectiveness loss, and model update consistency loss. Through data reconstruction, stealthy poisoned samples are generated, and effective poisoning attacks are achieved under a black-box assumption. Experiments conducted in both centralized and federated learning scenarios, across two datasets and four target models, demonstrate that CPDR reduces the model's classification accuracy on target classes with minimal impact on the performance of the main task. In centralized learning, CPDR achieves an additional reduction of 2.33 and 2.42 percentage points in target class accuracy compared to VagueGAN and Adversarial Poison, respectively. In federated learning, this advantage further increases to 4.44 and 4.10 percentage points, respectively. Moreover, the poisoned samples generated by CPDR exhibit minimal visual differences from the original samples, making them difficult to distinguish and detect, thus achieving a good balance between stealthiness and attack effectiveness.

Key words: data poisoning; poisoning attack; clean-label attack; AI security

1 引 言

1.1 研究背景

随着算法理论的创新和计算硬件的发展, 人工智能已被广泛应用于自动驾驶[1]、金融交易[2]、医疗诊断[3]、智能对话[4]等关键领域, 在推动生产力发展和改善个人生活方面发挥了重要作用. 深度学习是人工智能的核心技术之一, 其利用神经网络模型对数据进行特征提取和关系建模, 从而实现对复杂问题的预测和决策. 然而, 深度学习的安全性问题是当今备受关注的难题, 无论是传统的集中式学习还是新兴的联邦学习, 都面临诸多安全威胁. 近年来, 关于深度学习安全威胁的研究不断丰富和深入, 表明恶意用户能够利用模型反转攻击[5]、成员/属性推理攻击[6]、对抗样本攻击[7]以及投毒攻击[8]等多种攻击手段, 侵犯数据隐私或破坏模型可用性. 其中, 投毒攻击因其隐蔽性与破坏性, 已成为最具挑战性的威胁之一, 吸引众多学者研究其攻击机制.

数据投毒攻击是投毒攻击的形式之一, 指攻击者制备不利于模型训练的样本, 并用其污染模型训练集, 从而间接破坏模型. 数据投毒攻击具有现实意义, 因为深度学习模型的训练数据往往不完全可控: 一方面, 模型训练者可能通过各类众包平台收集数据, 但这些数据来源不明且缺乏严格审查, 参与者的匿名性和多样性导致恶意数据容易混入; 另一方面, 模型训练者还可能故意使用恶意数据进行训练, 例如在涉及利益冲突的联邦学习场景中, 恶意参与方可能故意发动攻击以阻止模型收敛.

数据投毒攻击可进一步分为脏标签攻击 (dirty-label attack) 和净标签攻击 (clean-label attack) 两类[9]. 脏标签攻击篡改训练样本标签, 引入错误的"特征-标签"映射关系, 致使模型学习到错误信息从而导致其产生预测偏差. 这类攻击容易实施且效果显著, 但因其标签篡改行为易被检测, 隐蔽性严重不足. 相比之下, 净标签攻击则通常基于对训练样本特征的扰动或直接生成, 能够在不改变标签的前提下, 诱导模型建立错误的"特征-标签"映射关系. 这类攻击中样本的视觉变化不明显, 难以被辨别和防御, 具有更强的隐蔽性和现实意义, 但对攻击者的技术水平要求较高, 且现有方案仍存在一定的局限性.

具体而言, 净标签攻击的实现方式主要包括两类. 第 1 类方式基于优化问题求解扰动数值, 通常依赖较强的白盒假设 (例如对目标模型梯度等信息的访问), 并通过多次迭代求解最优扰动以生成毒化样本. 这类方案在理论上具有直接和明确的优化目标, 但由于难以大规模并行制备毒化样本, 在实际部署中往往面临较高的计算开销, 且对目标模型或数据分布变化较为敏感. 相比之下, 第 2 类方式基于生成对抗网络 (generative adversarial network, GAN) 直接生成特征, 通常能够在黑盒假设下批量生成毒化样本, 且具备更强的泛化能力. 然而, GAN 的对抗训练机制容易出现模式崩塌等问题, 导致训练不稳定, 生成的样本质量参差不齐, 尤其在处理高分辨率图像时效果不佳.

针对上述两类净标签攻击方案的局限性, 本文提出一种基于数据重建的新型净标签数据投毒攻击方案, 旨在提升攻击的隐蔽性、有效性和鲁棒性. 本方案通过端到端的重建模型, 在特征空间中显示建模由"原始类"到"目标

类”的可控迁移, 从而生成具备较强隐蔽性的毒化样本. 本方案利用数据重建的单目最小化 (而非极大极小博弈) 优化目标提高训练过程的稳定性, 利用神经网络能够大规模并行生成样本的优势提高毒化样本制备效率, 且完全支持黑盒假设下的攻击场景. 本方案和上述两类方案的区别如表 1 所示.

表 1　3 类净标签毒化数据生成方案的对比

攻击方案	条件	优化目标	核心模型	稳定性	生成效率
基于优化问题	白盒	视具体方案	—	较稳定	批量生成受限
基于生成对抗网络	黑盒	极大极小博弈	生成器+判别器	易波动	离线训练+快速批量生成
基于数据重建	黑盒	单目标最小化	端到端重建模型	较稳定	离线训练+快速批量生成

本文的核心贡献可总结如下: (1) 将端到端的数据重建范式引入深度学习净标签投毒攻击问题, 提出新型攻击方案 CPDR, 通过数据重建生成具有较强隐蔽性的毒化样本以发动投毒攻击. (2) 基于 Attention U-Net 网络, 构建包含重建损失、投毒效果损失和模型更新一致性损失的多任务学习重建模型. (3) 在集中式学习和联邦学习两种攻击场景下, 利用多个数据集和目标模型进行实验, 并与 VagueGAN 和 Adversarial Poison 攻击方案进行对比, 证明了 CPDR 方案的有效性.

1.2　相关工作

数据投毒攻击的研究演进主要经历了从脏标签攻击向更具隐蔽性的净标签攻击的转变. 脏标签攻击的相关研究起源于传统机器学习模型. Biggio 等人[8]首次提出对传统支持向量机模型的投毒攻击, 其通过翻转部分训练样本标签成功干扰了支持向量机的决策, 该研究为以标签翻转为核心的脏标签攻击奠定了基础. 随后, 脏标签攻击从传统机器学习领域拓展到集中式深度学习领域. Liu 等人[10]通过逆向生成技术制备带有触发器的投毒样本, 指定其标签并用其对神经网络模型进行后训练, 从而使带有触发器的输入被误分类为攻击者指定的类别. 随着联邦学习的兴起, 对脏标签攻击的研究进一步扩展至分布式场景. Bagdasaryan 等人[11]指出联邦学习因其分布式特性更易受到投毒攻击, Tolpegin 等人[12]通过实验证明了标签翻转攻击在联邦学习中的有效性, 并进一步探究了攻击效果与恶意客户端比例和攻击时机等因素的关系. 为了增强攻击强度, Xiao 等人[13]研究了工业物联网环境中的联邦学习合谋投毒攻击, 其通过 Sybil 技术虚拟化出多个恶意节点, 并控制它们联合发动有效的标签翻转攻击. Gupta 等人[14]在标签翻转的基础上引入反转损失函数机制, 生成与模型正常更新方向偏离较大的恶意更新, 显著增强了攻击效果. 王瑞锦等人[15]提出了一种针对联邦原型学习 (FedProto) 的特征图投毒攻击方案, 通过标签翻转间接生成恶意特征图, 证明了基于特征聚合的联邦学习仍会受到投毒攻击影响. 然而, 脏标签攻击中样本特征与标签存在明显的不匹配现象, 导致其极易被数据审查和标签验证等机制检测, 或是被认证防御等机制在理论上限制其对模型影响的上界[16]. 随着针对脏标签攻击的防御技术不断成熟, 学术界逐渐将研究重心转向隐蔽性更强的净标签攻击.

净标签攻击要求投毒样本在视觉上与对应标签保持一致, 最初同样起源于集中式学习场景, 并逐渐被拓展到分布式场景, 其主流方案包括基于优化问题的特征扰动计算和基于 GAN 的直接特征生成. Shafahi 等人[17]首次提出净标签攻击, 通过前向后向分裂迭代算法对基样本进行微调, 使其特征接近目标类别样本而视觉表现不发生明显变化, 再将基样本加入训练集中引导模型定向出错. Fowl 等人[18]研究了原用于对抗样本攻击的对抗性样本, 指出对抗性样本也能作为净标签毒化样本以发动有效的投毒攻击. Zhao 等人[19]研究了深度学习中普遍存在的“天然毒化样本 (natural poisoned data)”, 利用精心构造的损失函数促使 GAN 生成与天然毒化样本分布接近的净标签毒化样本. Koh 等人[20]通过在样本特征空间建模并求解优化问题, 同样生成了与正常样本高度相似的净标签毒化样本, 成功绕过多种数据清理防御机制. Zeng 等人[21]研究了有限信息下的净标签攻击, 通过引入代理模型并迭代求解触发器, 实现了极低污染率下的高效攻击. 随着联邦学习的兴起, 研究者同样在联邦学习场景下成功实现了净标签投毒攻击. Yang 等人[22]在联邦学习场景下将攻击建模为优化问题并求解, 成功生成具有多样性的毒化样本, 能够绕过常见的鲁棒性聚合防御机制. Sun 等人[23]通过修改条件 GAN 的损失函数, 打破生成器与判别器的纳什均衡, 生成带有模糊效果的净标签毒化样本, 在联邦学习中实现了攻击效果和隐蔽性的平衡. Alharbi 等人[24]提出一种共谋后门攻击, 控制多个客户端注入隐蔽触发器并对毒化模型参数向量进行扰动调整, 同样保证了攻击效果与

隐蔽性. 尽管净标签攻击显著增强了投毒攻击的隐蔽性, 但现有方案仍存在一定局限. 基于优化问题的方案通常依赖现实中难以达成的白盒假设 (如模型梯度访问权限), 且串行的迭代计算限制了攻击效率; 而基于 GAN 的方案虽然实现了黑盒条件下的批量毒化样本生成, 但受限于对抗训练的不稳定性, 其在高维数据空间中的表征能力显著下降. 因此, 这两类方案均未同时满足实际场景下对攻击效率、隐蔽性和泛化性的要求, 具有进一步的提升空间.

2 基础知识

本节就本文提出的 CPDR 方案所涉及的相关概念和基本知识进行介绍.

2.1 净标签数据投毒攻击

净标签数据投毒攻击指攻击者在不篡改训练样本标签的情况下, 通过微调输入特征 (例如对图片像素进行微小修改) 或直接生成数据的方法得到毒化样本, 再将这些毒化样本注入深度学习模型的训练集中, 从而引导模型产生错误预测. 图 1 是净标签数据投毒攻击的一种简单示意.

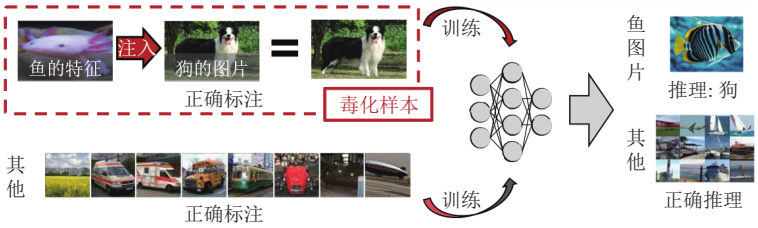


图 1 净标签数据投毒攻击示意图

如图 1 所示, 以毒蛙攻击 (poison frogs)[17]方案为例, 攻击者使用基于特征碰撞的方法, 微调图像像素, 使部分表示“狗”的图像的特征空间向“鱼”的特征空间偏移, 从而制成毒化样本. 毒化样本会影响模型的决策边界, 进而大概率将表示“鱼”的图像分类为“狗”. 联邦学习中的净标签投毒攻击与集中式学习场景相似, 不同点仅在于攻击者通常被假定为联邦学习参与方, 并将毒化样本注入本地训练集中, 从而间接污染本地模型, 再最终影响全局模型.

净标签投毒攻击具有以下特点: (1) 较强的隐蔽性: 攻击者无需篡改训练样本标签, 而是修改其特征, 能够绕过基于标签一致性的传统检测方法; (2) 特征依赖: 攻击方式主要为修改样本的特征 (如图像像素、文本嵌入向量等), 以引导模型学习错误的决策边界; (3) 定制化设计: 攻击者通常需要针对特定任务或特定类别精心设计毒化样本. 这些特点使净标签攻击相较于脏标签攻击更难被检测, 因为模型训练者难以察觉样本特征的微小变化, 难以检测标签一致性; 但与此同时, 净标签攻击也要求攻击者对目标模型架构或目标任务有深入了解, 并能据此进行复杂的毒化样本制备, 攻击发动难度高.

2.2 U-Net 网络模型

U-Net 神经网络模型由 Ronneberger 等人[25]提出, 是一种专为生物医学影像分割设计的深度学习模型. U-Net 模型的设计理念基于全卷积神经网络 (fully convolutional network, FCN) , 其主要特点是采用了对称的编码器-解码器结构, 并通过跳跃连接机制 (skip connections) 将编码器和解码器的特征图连接起来, 从而保留图像中的细节信息, 避免信息损失. 该模型最初用于生物医学影像的语义分割任务, 但其也因强大的特征提取和恢复能力, 逐渐被用于图像生成任务中, 例如被用于潜空间扩散模型 (latent diffusion model, LDM) 作为底层像素生成器.

如图 2 所示, U-Net 模型采用对称的 U 形结构 (图中将其旋转以节约空间) : 上侧是编码器部分, 通过一系列卷积和池化操作逐渐提取图像的高阶特征; 下侧是解码器部分, 通过反卷积上采样和卷积操作恢复图像的空间分辨率. 跳跃连接使浅层特征 (如边缘和纹理信息) 可以直接传递到解码器中, 有效融合浅层细节信息与深层语义特征, 实现高效信息传递, 从而在重建图像时保持更高的细节质量. 此外, 这种特征重用和多层信息融合的机制也有利于 U-Net 模型适应小样本数据集.

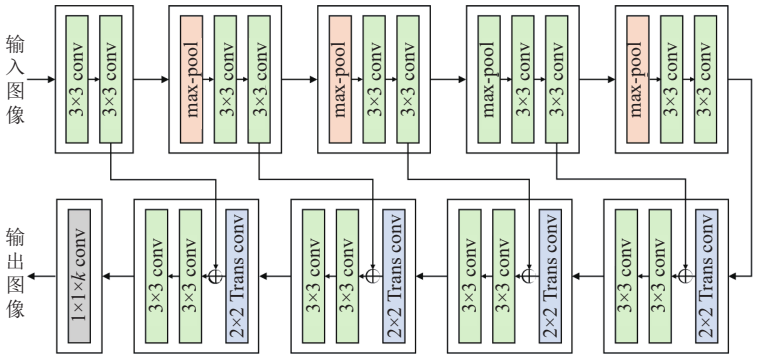


图 2 U-Net 模型结构

3 基于数据重建的净标签数据投毒攻击方案 CPDR

为进一步提高净标签数据投毒攻击的隐蔽性和有效性, 本文提出一种基于数据重建的净标签数据投毒攻击 (clean-label data poisoning attack via data reconstruction, CPDR) 方案, 该方案将正常样本重建为毒化样本并混入训练集, 引导目标模型对特定类别样本分类出错. 本节将首先确立系统模型与威胁模型, 明确攻击场景与前提假设, 随后详细介绍该方案关键模块的设计原理与实现机制, 最后系统地总结出完整攻击流程.

3.1 系统模型和威胁模型

CPDR 方案在集中式学习和联邦学习场景中均可应用, 如图 3 所示. 在集中式学习中, 模型训练者 T 通过多源数据 D_i 训练一个深度学习模型. 为干扰训练过程, 攻击者 A 在其持有的正常数据 D_{atk} 上使用 CPDR 方案生成一系列毒化数据 D_p, 并将其注入 T 的训练集中. 这种注入可以通过渗透数据供应链、利用数据众包平台或污染自动化数据采集管道等方式实现. 而在联邦学习中, 聚合服务器 S 协调多个参与方 C_i 共同训练全局模型. 此时, 攻击者 A 生成毒化数据 D_p 后, 还需将 D_p 混入本地模型的训练集中, 从而训练出恶意的本地模型并间接破坏全局模型.

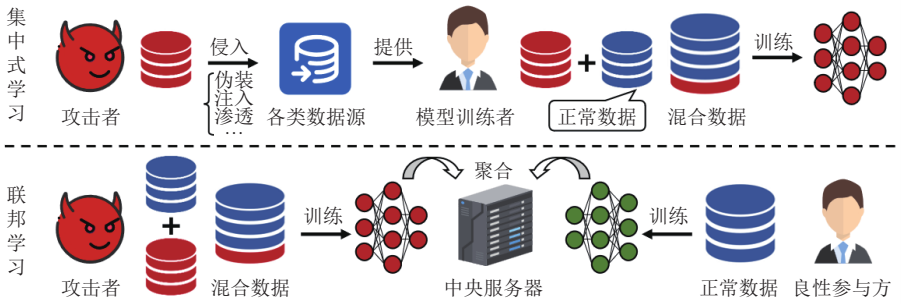


图 3 系统模型

对于威胁模型, 本文从攻击者的目标、知识和能力这 3 个方面做出假设.

• 攻击者目标: 与其他相关研究[17,19,21,23]类似, 本文也主要关注图像数据和图像分类模型. 作为一种净标签定向数据投毒攻击, 攻击者的目标是生成视觉表现与标签一致的净标签毒化样本, 促使模型倾向于对特定类别样本分类出错. 在现实场景下, 这类攻击通常出于经济动机, 例如通过污染对手的训练数据使其模型能力下降 (如内容推荐准确率、视觉识别准确率), 从而在商业竞争中取得优势, 或对金融风控及预测模型实施数据干扰以获取非法收益.

• 攻击者知识: 本文假设攻击者具有有限的知识, 符合黑盒假设. 具体而言, 假设攻击者具有正常的本地数据集, 且在联邦学习下进一步假设其数据集与其他参与方分布相似. 然而, 攻击者不能直接访问目标模型, 因此无法利用目标模型训练时的梯度和损失函数值等信息.

• 攻击者能力: 本文假设攻击者能够完全控制其本地数据集, 例如基于本地数据进行修改、重建等操作. 在集中式学习中, 攻击者充当数据提供者将毒化样本注入目标模型训练集中; 而在联邦学习中, 攻击者直接以参与方的身份加入联邦学习流程, 并使用制备的毒化样本执行本地训练.

3.2 整体设计

根据第 3.1 节的系统模型和威胁模型, 攻击者在满足净标签约束的前提下, 仅能通过向训练数据中注入精心构造的样本来干扰模型. 因此, 如何在有限能力下制备兼具隐蔽性与攻击性的毒化样本, 成为实现数据投毒攻击的关键问题. 本节将围绕毒化样本的制备, 详细介绍 CPDR 攻击方案的总体思路和设计细节.

现有的数据投毒攻击方案主要分为两类: 基于优化问题的攻击方案与基于生成模型的攻击方案. 前者通过迭代计算精确的扰动值以最小化攻击目标函数, 因此通常依赖对白盒信息 (如目标模型梯度) 的访问并伴随较高的计算开销. 后者则通过对抗性训练生成毒化样本, 这类攻击虽支持更宽松和实际的假设 (黑盒假设), 但对抗训练范式普遍存在训练不稳定、模式崩塌与对精细语义控制能力不足的问题, 导致其生成样本的质量通常不如仅添加微小扰动的样本, 从而影响攻击的隐蔽性. 基于上述分析, 本文提出的 CPDR 方案摒弃传统的精确扰动计算或噪声直接生成的范式, 而是设计了以端到端数据重建为核心的新型毒化样本生成范式. 该范式的基本思想是: 利用具备强特征提取与像素级数据重建能力的模型, 在保持输入样本整体语义和像素结构高度一致的前提下, 引导其深层特征向预设方向发生系统性偏移, 从而实现“将正常样本重建为毒化样本”, 以替代复杂的优化计算或不稳定的对抗生成.

具体而言, CPDR 方案关注到 U-Net 模型在多尺度特征建模和数据重建任务中的优势, 通过引入特定的特征提取器并重新设计损失函数, 使模型在重建数据的同时, 显式地推动输出样本在深层特征空间中产生偏移. 这种新范式摆脱了对白盒假设的依赖, 同时避免了对训练数据整体分布的显式建模, 使毒化样本生成过程更加明确和稳定. 在保持样本整体语义一致的前提下, 该范式能够有效诱导输出样本的深层特征偏移, 从而满足净标签数据投毒攻击对隐蔽性和有效性的要求. 得益于上述设计, CPDR 方案摆脱了对目标模型内部参数的实时依赖, 支持在黑盒场景下批量化生成毒化样本, 又采用具有明确全局最小化目标的重建模型, 避免了对抗训练中常见的模式崩塌与梯度震荡等不稳定的问题. 通过在特征空间中显式推动特征偏移, 该方案在保持样本在人类视觉感知下的“净标签”特性的同时, 展现出在模型表征空间内的极强干扰能力.

攻击者使用 CPDR 方案重建出毒化样本集合后, 即可发动投毒攻击. 根据第 3.1 节的系统模型, 在集中式学习场景下, 攻击者可以作为数据提供方, 通过数据众包平台等渠道将毒化样本传播给模型训练者; 在联邦学习场景下, 攻击者可以作为参与方, 按照一定比例将毒化样本混入正常的本地数据集, 并利用混合数据集进行本地模型训练, 从而最终间接破坏全局模型. CPDR 方案的整体架构如后文图 4 所示. 除核心的数据重建模型外, 该方案还需引入深层特征提取模块并设计新的综合损失函数, 分别用于优化输出样本的重建质量、深层特征偏移程度以及其对全局模型训练结果的影响. 在这些优化目标的协同约束下, CPDR 方案能够生成兼具隐蔽性与攻击性的毒化样本集合. 下面将分别对这些模块进行详细介绍.

3.2.1 重建模型结构

CPDR 方案的核心即为重建模型, 本节将介绍其选型动机和具体模型结构. 现有的图像到图像 (image-to-image) 生成模型主要包括 3 类: GAN 变体 (如 Pix2Pix 模型)、基于概率建模的自编码结构 (如条件变分自编码器模型) 和扩散模型. 这些模型虽广泛应用于图像风格转换、超分辨率、照片修复等任务, 但应用于本文的数据重建场景时存在一定局限: Pix2Pix 等基于 GAN 的模型会受到对抗性训练带来的模式崩塌影响; 条件变分自编码器等基于概率建模的模型则对训练数据的分布有先验假设, 其生成图像的细节质量通常表现一般; 基于扩散模型的方案则通常从噪声生成图像, 输入样本只能作为条件输入, 面临较大的计算资源和计算效率压力, 且没有有效利用输入图像的信息.

综合考虑上述局限, CPDR 方案中的重建模型采用 Attention U-Net[26]结构. 该结构是一种原始 U-Net 的改进型结构, 专为图像处理任务设计, 尤其适用于生物医学影像分割. 其输入为固定尺寸的三维图像张量 (通道数×宽×

高), 经过模型处理后, 输出与输入图像尺寸相同的重建图像或分割掩码. 相较于前述生成模型, Attention U-Net 具有 3 方面的优势: 首先, 该结构无需对抗性训练, 规避了 GAN 的训练不稳定问题; 其次, 该结构不依赖于先验假设, 泛化性能更强; 最后, 该结构避免了扩散模型中扩散过程的多步迭代采样, 计算效率显著提升. 这些特性使得该结构更契合本文攻击场景对重建精度、训练稳定性及效率的综合需求.

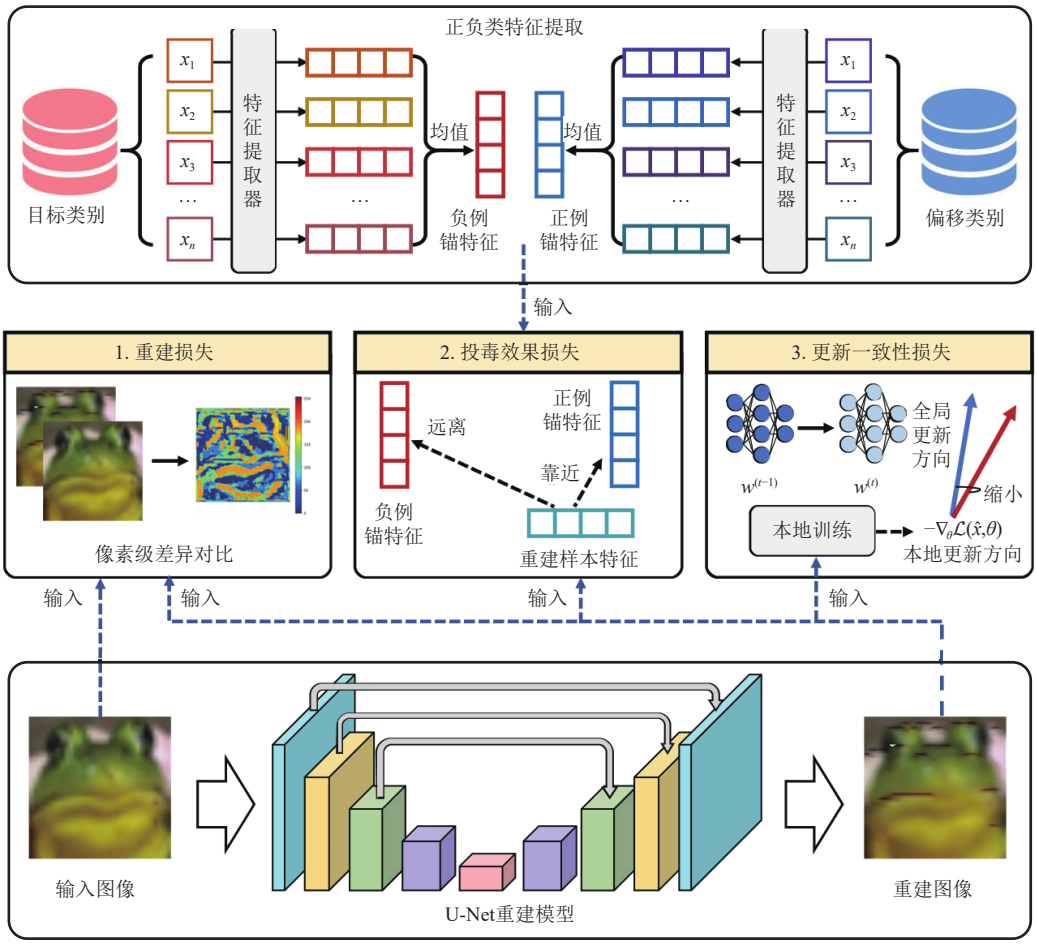


图 4 CPDR 方案架构

如图 5 所示, Attention U-Net 模型由编码器、解码器、跳跃连接和注意力模块这 4 部分组成. 编码器通过逐层卷积与下采样操作提取多尺度特征; 解码器则通过反卷积逐步恢复图像空间分辨率; 与传统的 U-Net 模型不同, Attention U-Net 模型在跳跃连接时引入注意力门控 (attention gate) 机制, 用于在融合编码器和解码器特征时动态筛选出与攻击目标相关的重要特征, 同时抑制不相关或冗余的信息.

注意力门控机制的处理过程主要分为 3 步.

(1) 它接收来自编码器的跳跃连接特征图 x 和来自解码器的门控信号 g, 二者通过线性变换 (例如 1×1 卷积) 被映射到一个统一空间:

$$x' = W_x * x + b_x,\ x \in \mathbb{R}^{B\times C_x\times H_x\times W_x} \tag{1}$$

$$g' = W_g * g + b_g,\ g \in \mathbb{R}^{B\times C_g\times H_g\times W_g} \tag{2}$$

其中, *表示卷积操作, H_x 和 W_x 通常大于 H_g 和 W_g, 即特征图 x 比门控信号尺寸更大.

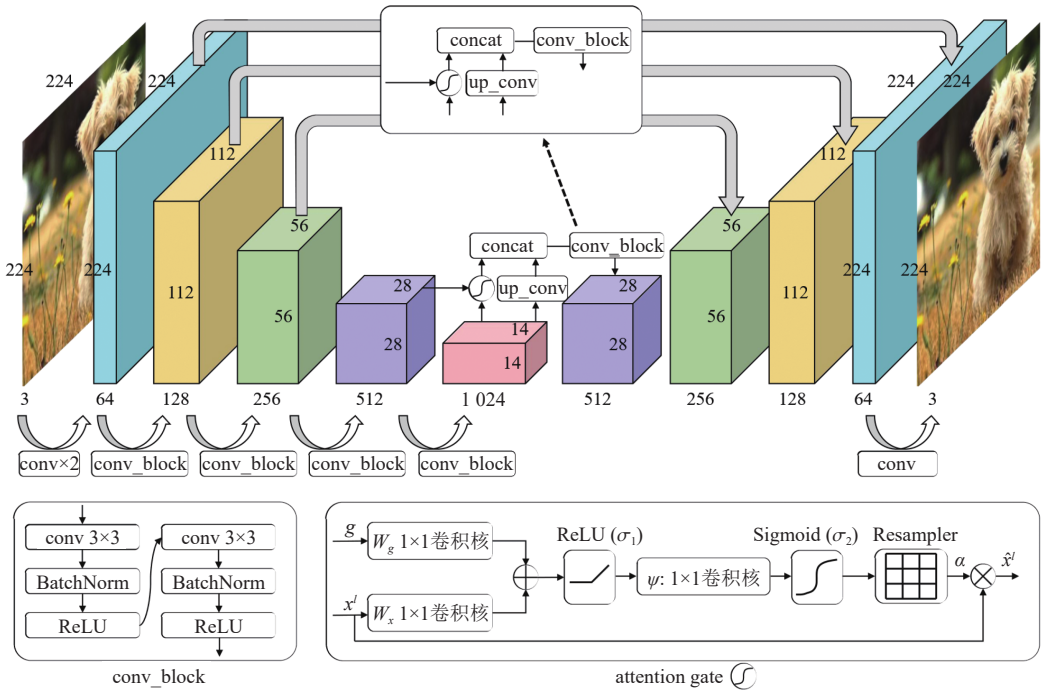


图 5 重建模型结构

(2) 特征图与上采样后的门控信号进行加和, 再经过非线性变换后得到注意力中间表示:

$$q = ReLU(x' + Upsample(g')) \tag{3}$$

随后使用一个 1×1 卷积核 ψ 将其映射为单通道的注意力响应, 再经 *Sigmoid* 激活函数归一化:

$$\alpha = Sigmoid(\psi * q + b_\psi) \tag{4}$$

此时得到的 α 即为注意力系数图, 表示输入的特征图中每个空间位置的重要性.

(3) 归一化后的注意力系数与编码器特征逐元素相乘, 得到按注意力筛选后的特征, 再与解码器特征进行融合:

$$\tilde{x} = \alpha \odot x \tag{5}$$

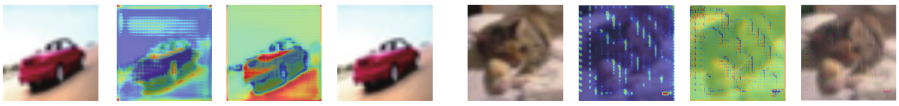
此时的 $\tilde{x}$ 就对应原始 U-Net 中的跳跃连接输出, 相当于在原本的跳跃连接特征图上增加了一次注意力加权. 这种机制使解码器接收到的特征图既保留编码结果的空间细节, 又含有和实际任务相关的动态调整, 有利于模型在数据重建任务中提高对关键区域的特征捕获能力, 抑制无关区域的影响.

这种优势在净标签毒化样本生成的任务中十分关键, 因为攻击者需要确保深层特征偏移方式能被该模型有效学习, 同时弱化无关特征的干扰. 此外, Attention U-Net 模型的结构对多尺度特征的捕捉能力较强, 能够实现复杂场景下视觉表现高度相似的图像重建, 提升毒化样本的隐蔽性. 综上, Attention U-Net 模型的特性非常适合作为净标签投毒攻击中的毒化样本生成器, 本文选用其作为重建模型, 以生成高质量的投毒样本.

本文对 attention gate 的效果进行了简单验证, 具体来讲, 取 Attention U-Net 网络中最后两个 attention gate 的中间输出注意力系数 α, 将其转换为热力图并叠加在输入图像 x 上, 观察其结果. 如图 6 所示, 图 6(a) 是仅使用重建损失项的结果, 注意力集中于输入样本像素分布相对复杂的区域 (颜色更多样的区域), 关注这些复杂区域的重建; 而如果加入投毒效果损失, 注意力会集中在一些特殊的纹理上, 最终的重建图像隐约存在这些纹理, 正是通过这种对输入图像的精细微调, 才能最终促使重建模型深层特征改变.

需要指出的是, 重建模型结构只能保证生成样本和原始样本在像素空间上的一致性, 其本身并不能直接约束

输出样本在深层特征空间中的偏移方向与程度. 这种特征偏移依赖专门的深层特征提取器, 并通过相应的损失函数约束将其纳入整体框架. 第 3.2.2 节将对该深层特征提取器进行介绍.



(a) 图像重建　　(b) 图像重建+特征偏移

图 6　Attention gate 输出示意图

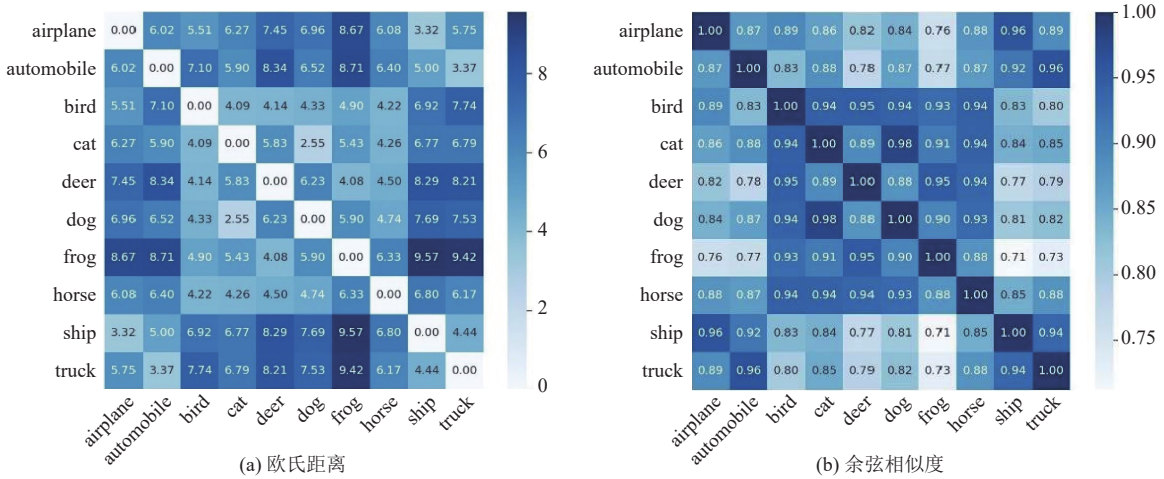
3.2.2　深层特征提取器

本节将介绍 CPDR 方案采用的深层特征提取器, 及其在损失函数构建中的作用. 目前已有许多关于特征提取的研究, 例如主成分分析、滤波器组以及更复杂的基于图模型或元学习的方法等. 这些方案在特定场景中具有一定的应用价值, 但在图像处理方面仍然存在局限. 研究表明[27], 经过大规模预训练的卷积神经网络模型或其变体, 配合简单的多层感知机分类头, 即可实现小样本图像分类, 且效果优于精心设计的基于元学习等技术的特征提取器. 这说明预训练卷积神经网络具备高效且泛化性强的深层特征提取能力. 因此, 本文采用这种基于神经网络的特征提取方式, 取神经网络最后一个隐藏层的输出作为样本深层特征.

为促使重建毒化样本的深层特征向特定类别偏移, 需提取每一类别具有较强判别性的代表特征作为锚特征. Papyan 等人[28]和 Rangamani 等人[29]的研究都表明, 在深度神经网络训练的后期阶段, 属于同一类别的样本深层特征会自然地收敛到其类别中心 (即平均特征), 且各个类别中心会非常均衡地分散开, 趋向于一种称为“等角单纯形紧帧 (simplex equiangular tight frame, ETF)” 的对称几何结构. 这一结论说明分类任务中不同类别样本的平均特征能够作为该类有效的判别性特征. 因此, 本文也采用每类样本的平均特征作为该类的锚特征, 用于在损失函数中引导重建样本的特征偏移方向.

为衡量重建样本相对于目标类别的特征偏移程度, 需要确定量化特征差异的指标. 考虑到欧氏距离能够同时反映特征向量的方向与模长差异, 更适合衡量类别间深层特征的分布差距, 且已有研究 (如 FaceNet[30]、Prototypical Networks[31]) 证实了其在高维特征空间中的有效性, 本文采用欧氏距离作为特征差异度量指标.

本节以 ResNet-18 模型为例, 使用去除分类头的预训练 ResNet-18 模型对 Cifar10 数据集进行特征提取. 如图 7 所示, 不同类别样本特征均值的欧氏距离差异显著, 而余弦相似度差异则不明显, 验证了上述选型决策的合理性. 综上, CPDR 方案最终采用去除分类头的预训练卷积神经网络模型充当特征提取器 (根据数据集特点进行具体选择, 本文最终采用 EfficientNet-B0 和 ResNet-50 模型), 以类别平均特征为锚特征, 以欧氏距离衡量不同类别之间的深层特征差异. 上述特征提取与度量策略将作为特征约束项被引入损失函数中, 引导重建样本的深层特征偏移方向与程度.



(a) 欧氏距离　　(b) 余弦相似度

图 7　Cifar10 数据集各类别特征均值差异

3.2.3 损失函数设计

在上述重建模型结构与深层特征提取策略的基础上, 本节进一步设计综合损失函数, 对毒化样本的生成过程进行联合约束与优化.

Attention U-Net 模型原本用于医学影像分割任务, 其输出为与输入图像尺寸一致的概率热图. 概率热图是一个元素取值范围在 0 到 1 之间的矩阵, 并将被进一步二值化为只包含 0 或 1 的矩阵, 0 表示该像素位置被模型判定为背景部分, 1 表示该像素位置被模型判定为主体部分 (如肿瘤块). 用于医学影像分割时, Attention U-Net 原始的损失函数为像素级二元交叉熵损失 (binary cross entropy loss, BCE Loss), 即把对每个像素的分类视为一个二分类任务 (判断该像素属于背景或主体), 再整体进行平均值计算, 具体表示如下:

$$L_{\mathrm{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)] \tag{6}$$

其中, y_i 是第 i 个像素的真实标签, $\hat{y}_i$ 是第 i 个像素被预测的标签, N 是像素总数.

此外, 常用的损失函数还有骰子损失 (Dice Loss), 其衡量被分割出的图像主体部分与真实的主体部分重叠的程度, 具体表示如下:

$$L_{\mathrm{Dice}} = 1 - \frac{2|X \cap Y|}{|X| + |Y|} \tag{7}$$

其中, X 表示真实主体部分所含的像素数量, Y 表示模型分割出的主体部分所含的像素数量, $|X \cap Y|$ 则表示二者交集部分的像素数量.

在 CPDR 方案中, Attention U-Net 模型作为图像重建模型, 任务目标发生变化, 其输出的含义不再是概率热图, 因此上述损失函数也不再适用, 需要设计新的损失函数. CPDR 方案最核心的任务是毒化样本生成, 即攻击者需要训练出一个能把正常样本重建为毒化样本的最佳重建模型. 本文认为, 重建模型需要满足以下条件: 第一, 视觉隐蔽性. 在净标签的前提下, 毒化样本在视觉表现上应尽可能接近正常样本; 第二, 投毒有效性. 重建模型需要引导毒化样本深层特征偏移, 使目标模型学习到错误的“特征-标签”映射, 进而造成其决策边界变化并在推理过程产生错误; 第三, 更新一致性. 在联邦学习场景中, 攻击者还需要保证其上传的本地更新方向与其他参与方没有显著差异, 以避免其本地更新被中央服务器的鲁棒性聚合策略丢弃. 因此, 为同时实现上述 3 大目标, 本文设计出一个包含 3 部分的综合损失函数以取代 Attention U-Net 模型的原始损失函数. 具体来说, 综合损失函数由以下 3 部分组成.

首先是重建损失 L_{recon}, 该损失旨在确保毒化样本在视觉上与正常样本保持高度一致. 为实现该目标, 本研究将重建模型 F_{R} 的输入样本 x 和其重建样本 $\hat{x} = F_{\mathrm{R}}(x)$ 的像素均方误差 (mean squared error, MSE) 作为重建损失, 即:

$$L_{\mathrm{recon}} = \|x - F_{\mathrm{R}}(x)\|_2^2 \tag{8}$$

重建损失能够有效减少毒化样本在像素级别的异常, 使其难以被标签审查等传统机制检测.

其次是投毒效果损失 L_{atk}, 该损失旨在确保毒化样本发生深层特征偏移, 本文引入一种类似 Triplet Loss 的相对距离度量机制对其进行量化. Triplet Loss 是度量学习中常用的目标函数, 常用于训练编码模型 (embedding model), 它通过约束“锚点样本-正样本-负样本”三元组之间的相对距离, 使模型在特征空间中学习到判别性更强的嵌入表示, 且三元组对应的样本动态变化. 但在本方案中, 被固定为锚点的正样本和负样本, 分别是使用特征提取器 F_{E} 计算出的目标类别 (正类) 和原始类别 (负类) 所有样本的深层特征均值, 记为 $\bar{f}_{\mathrm{pos}}$ 和 $\bar{f}_{\mathrm{neg}}$. 在训练重建模型时, 提取毒化样本 $\hat{x}$ 的深层特征 $F_{\mathrm{E}}(\hat{x})$, 并构造三元组 $(F_{\mathrm{E}}(\hat{x}), \bar{f}_{\mathrm{neg}}, \bar{f}_{\mathrm{pos}})$, 按照公式 (9) 优化重建模型:

$$L_{\mathrm{atk}} = \max(0, d(F_{\mathrm{E}}(\hat{x}), \bar{f}_{\mathrm{pos}}) - d(F_{\mathrm{E}}(\hat{x}), \bar{f}_{\mathrm{neg}}) + margin) \tag{9}$$

其中, $d(\cdot, \cdot)$ 表示特征向量间的欧氏距离, $margin$ 为超参数, 用于控制目标类与原类在特征空间中的最小间隔. 当重建样本深层特征距正锚点的距离和距负锚点的距离差值大于 $margin$ 时, 该损失项才能为 0, 否则仍会有梯度信

号促使模型参数优化. 但需要注意的是, 该参数的选取通常与具体的数据集 (甚至类别) 及特征提取器结构有关, 本文主要通过实验反馈动态调节其取值, 通常取 0.5–1.0 之间.

该损失项的优化目标是使毒化样本的深层特征向量更接近目标类别特征均值 $\bar{f}_{\text{pos}}$, 同时远离原始类别特征均值 $\bar{f}_{\text{neg}}$, 从而诱导分类模型决策边界改变. 相比于均方误差等绝对距离的约束方式, Triplet Loss 的相对度量特性在优化过程中同时考虑“靠近目标类”与“远离原类”两个方向, 使梯度更新更加平衡与平滑, 从而更有效地实现重建样本的深层特征偏移, 增强攻击效果.

最后是模型更新一致性损失 L_{update}, 该损失仅用于有防御的联邦学习场景, 旨在限制本地模型更新方向与全局模型更新方向的偏差, 以绕过中央服务器的鲁棒性聚合策略, 从而进一步提高攻击隐蔽性. 在联邦学习场景中, 使用毒化样本 $\hat{x}$ 训练当前本地模型并得到本地更新 $w_{\text{atk}}^{(t)}$ 的过程可表示为:

$$w_{\text{atk}}^{(t+1)} = w^{(t)} - \gamma \frac{\partial L_{\text{global}}(\hat{x}, w^{(t)})}{\partial w^{(t)}} \tag{10}$$

其中, $w^{(t)}$ 表示本轮全局模型参数, L_{global} 是全局模型损失函数, γ 是学习率数值. 因此, 对于攻击者而言, 在毒化样本的作用下, 其本地更新方向可以表示为$-\partial L_{\text{global}}(\hat{x}, w^{(t)})/\partial w^{(t)}$. 而对于良性参与方而言, 中央服务器每轮都会向参与方发送全局模型参数, 本轮和上一轮全局模型参数的差异 $w^{(t)}-w^{(t-1)}$ 是良性参与方共同作用的结果, 其可以作为全局模型的正常更新方向. 因此, 只需要使得攻击者本地更新方向尽可能接近全局模型的正常更新方向, 即可达成全局模型更新一致性. 因此, 模型更新一致性损失可以表示为:

$$L_{\text{update}} = 1 - Cosine\_Similarity\left(-\frac{\partial L_{\text{global}}(\hat{x}, w^{(t)})}{\partial w^{(t)}}, w^{(t)} - w^{(t-1)}\right) \tag{11}$$

综上所述, 综合损失函数包含 L_{recon}、L_{atk} 和 L_{update} 这 3 部分 (集中式学习场景中忽略 L_{update}, 下文不再赘述), 攻击者需要同时优化这 3 个目标函数, 以确保毒化样本在视觉上难以被察觉、深层特征发生偏移, 且受其影响的本地更新与全局模型更新方向接近. 这一过程可以视为一个典型的多任务学习问题, 因此需要为每个损失函数赋予权重参数以实现不同目标之间的平衡, 即:

$$L_{\text{total}} = \lambda_1 L_{\text{recon}} + \lambda_2 L_{\text{atk}} + \lambda_3 L_{\text{update}} \tag{12}$$

在多任务学习中, 平衡不同任务的损失权重一直是关键问题. 实践中可通过训练反馈手动微调, 或逐个任务调参来应对常规情况. 如果模型难以收敛, 可以进一步使用 PCGrad、Uncertainty Weighting 等智能算法, 这些智能算法支持自动分析各任务的梯度冲突或不确定性, 从而动态调整权重以提升训练效率.

3.3 攻击流程

在第 3.2 节中, 本文从模型结构、特征提取和损失函数设计这 3 个方面介绍了 CPDR 方案的核心模块, 本节将进一步说明攻击者如何将上述模块结合以制备毒化样本并发动攻击. CPDR 方案主要包含毒化样本制备和投毒攻击两个主要流程. 假设攻击者希望尽可能使类别 A 的样本深层特征偏移向类别 B, 则需要按照算法 1 给出的流程进行毒化样本制备.

算法 1. 基于数据重建的毒化样本生成.

输入: 攻击者数据集 $D_{\text{atk}} = \{(x_i, y_i)\}$; 重建模型 F_{R}; 特征提取器 F_{E}; 全局模型参数 $w^{(t)}$, $w^{(t-1)}$; 学习率 η; 损失函数超参数 λ_1, λ_2, λ_3; 批处理大小 b; 投毒效果损失超参数 *margin*; 重建模型训练轮数 T;
输出: 毒化样本集 D_p.

1. 初始化重建模型 F_{R} 的参数 θ
2. $D_A \leftarrow \{x_i \mid x_i \in D_{\text{atk}} \text{ and } y_i = A\}$
3. $\bar{f}_{\text{pos}} \leftarrow \text{Avg}(F_{\text{E}}(x_i) \mid y_i = B).\text{repeat}(b, 1)$ // 提取目标类别特征向量
4. $\bar{f}_{\text{neg}} \leftarrow \text{Avg}(F_{\text{E}}(x_i) \mid x_i \in D_A).\text{repeat}(b, 1)$ // 提取原始类别特征向量
5. $w_{\text{atk}}^{(t)} \leftarrow w^{(t)}$ // 接收全局模型参数

6. **FOR** $t_r \in (1, 2, \ldots, T)$ **DO**
7. **FOR** $mini\_batch\{(x_i, y_i)\}_{i\in(1, 2, \ldots, b)} \in D_A$ **DO**
8. // (1) 重建样本
9. $x \leftarrow mini\_batch\{x_i\}_{i\in(1, 2, \ldots, b)}$
10. $\hat{x} \leftarrow F_R(x)$ // 使用重建模型生成重建样本
11. // (2) 损失计算
12. $L_{recon} \leftarrow \|x - \hat{x}\|_2^2$ // 计算重建损失
13. $f_{\hat{x}} \leftarrow F_E(\hat{x})$ / 提取重建样本的深层特征
14. $L_{atk} = \max(0, d(f_{\hat{x}}, \bar{f}_{pos}) - d(f_{\hat{x}}, \bar{f}_{neg}) + margin)$ // 计算投毒效果损失
15. $\Delta w_{atk} \leftarrow -\frac{\partial L_{global}(\hat{x}, w^{(t)})}{\partial w^{(t)}}$ // 计算本地更新方向
16. $\Delta w_{global} \leftarrow w^{(t)} - w^{(t-1)}$ // 计算全局模型的正常更新方向
17. $L_{update} \leftarrow 1 - Cosine\_Similarity(\Delta w_{global}, \Delta w_{atk})$ // 计算模型更新一致性损失
18. $L_{total} = \lambda_1 L_{recon} + \lambda_2 L_{atk} + \lambda_3 L_{update}$ // 计算综合损失
19. // (3) 参数更新
20. $\theta \leftarrow \theta - \eta \frac{\partial L_{total}}{\partial \theta}$ // 使用梯度下降法更新参数
21. **END FOR**
22. **END FOR**
23. // (4) 毒化样本生成
24. $num\_poi \leftarrow |D_{atk}| \times poi\_rate$ // 采样固定比例正常样本
25. $mini\_batch\{(x_i, y_i)\}_{i\in(1, 2, \ldots, num\_poi)} \leftarrow Random\_Sample(D_A, num\_poi)$
26. $x \leftarrow mini\_batch\{x_i\}_{i\in(1, 2, \ldots, num\_poi)}$
27. $x_{poi} \leftarrow F_R(x)$ //毒化样本生成
28. $D_p \leftarrow \{x_i \mid x_i \in x_{poi}, i \in (1, 2, \ldots, num\_poi)\}$
29. **RETURN** D_p

(1) 初始化: 攻击者初始化用于生成毒化样本的重建模型 F_R 的参数 θ, 并从自己的数据集中筛选出所有类别 A 的样本构成数据子集 D_A. 随后, 通过特征提取器 F_E 提取目标类别 B 中全部样本的深层特征向量, 计算其均值 $\bar{f}_{pos}$, 并将该特征向量复制扩展为含 b 个相同向量的批量特征向量, 用于适配后续的批量运算. 使用同样的方法从类别 A 中提取 $\bar{f}_{neg}$.

(2) 重建模型训练: 重建模型分 T 轮进行训练, 每轮基于类别 A 的样本子集 D_A 按照固定的批量大小 b 迭代处理. 在每次迭代中, 首先将批量样本 x 输入至重建模型 F_R 中以生成对应的重建样本 $\hat{x}$. 然后, 根据重建样本和其他已有信息, 依次计算重建损失、投毒效果损失和模型更新一致性损失:

• 重建损失 L_{recon}: 衡量重建样本与原始样本之间的差异, 使用重建样本与原始样本像素级均方误差计算, 确保生成的毒化样本视觉上与原始样本接近, 确保视觉隐蔽性.

• 投毒效果损失 L_{atk}: 衡量重建样本相比于原始样本的深层特征偏移程度, 以 $\bar{f}_{pos}$ 和 $\bar{f}_{neg}$ 分别作为正负锚特征, 计算重建样本的特征向量和正负锚特征构成的 Triplet 损失, 确保重建样本的深层特征成功偏移.

• 模型更新一致性损失 L_{update}: 通过余弦相似度约束毒化样本训练出的本地更新方向与全局正常更新方向的一致性, 规避异常检测. 此损失函数仅在有防御的联邦学习中使用, 在集中式学习中不涉及该部分损失.

综合损失 L_{total} 通过加权求和上述 3 项损失得到, 使用梯度下降法对重建模型参数 θ 进行更新.

(3) 毒化样本生成: 在重建模型训练完成后, 攻击者按照一定比例从类别 A 的样本子集中随机采样, 生成一批毒化样本. 通过重建模型对这些样本进行转换, 生成最终的毒化样本集 D_p.

(4) 投毒攻击发动: 制成毒化样本数据集后, 攻击者需要发动投毒攻击. 在集中式学习中, 攻击者不参与目标模型训练, 但需要把其生成的毒化样本提供给模型训练者. 如果模型训练者通过开放数据众包平台 (如 Amazon Mechanical Turk) 收集数据, 攻击者可以伪装成合法数据贡献者提交毒化样本集 D_p. 如果模型训练者依赖第三方数据提供商 (如医疗影像标注公司、传感器数据提供商), 攻击者可以通过渗透或收买供应商内部人员, 将 D_p 混入合法数据流. 如果模型训练者采用自动化数据采集管道 (如实时爬虫、用户上传接口), 攻击者可以通过在采集源高频提交毒化数据, 使 D_p 被系统自动收录. 以上投毒策略非常灵活, 本文不讨论其具体实现. 而在联邦学习中, 投毒策略相对固定. 攻击者作为联邦学习的参与方, 直接使用 D_p 训练自己的本地模型, 并将训练后本地模型的参数上传即可. 需要强调的是, 根据文献 [12] 的研究, 在联邦学习场景下, 数据投毒攻击者并非每轮都需要发动攻击, 因为在初始几轮中全局模型的更新并不稳定, 攻击者仅在最后几轮投毒即可达到较好的效果. 因此, 毒化样本制备与投毒攻击在时序上分离, 即攻击者在制备阶段扮演良性参与方正常执行训练协议, 待毒化样本制备完毕后, 攻击者在训练的最后几轮执行投毒攻击.

4 实验分析

本节通过实验对 CPDR 方案进行效果评估和分析, 并与近几年的其他方案进行对比. 大量实验结果表明, 在集中式学习场景和联邦学习场景中, 该方案都能够隐蔽和有效地实现定向投毒攻击.

4.1 实验设置

本文实验在云服务器上进行, 该服务器配备 Intel(R) Xeon(R) Platinum 8474C CPU 和 RTX 4090D (24 GB) GPU, 使用 Ubuntu 22.04 操作系统和 PyTorch 2.1.2 (CUDA 11.8) 框架. 本文实验使用 FashionMNIST 和 Cifar10 两个经典基准数据集. 前者包含 70 000 张 28×28 的灰度图像, 后者包含 60 000 张 32×32 的 RGB 彩色图像. 目标模型选用 LeNet-5、Mini-VGG、MobileNet-v2 和 ResNet-18 这 4 种具有代表性且被广泛使用的模型. 下文将首先展示 CPDR 生成的毒化样本的视觉表现, 以表明其隐蔽性, 随后介绍对其攻击能力的评价指标, 并分别在集中式学习和联邦学习环境下展开评估, 以表明其攻击有效性. 相关的训练超参数和投毒比例将在相应的实验部分详细说明. 本方案的实现代码可见 https://github.com/CPDR2026/CPDR.

4.2 毒化样本质量

图 8 和图 9 分别展示了 CPDR 方案在 FashionMNIST 和 Cifar10 数据集上生成的毒化样本的视觉效果. 对于每类图像, 展示出 3 组具有代表性的样本对, 每组包含一个作为输入的原始正常样本及其对应的重建样本. 在重建过程中, 重建模型在基本保持输入样本像素的情况下, 向输入样本添加特定噪声, 使输出的重建样本能够成功误导分类模型. 结果表明, CPDR 方案具有强大的样本重建能力, 能够为任意给定的正常样本生成对应的毒化版本. 从视觉质量来看, 这些毒化样本保持了较高的保真度, 仅包含微弱的噪声痕迹, 这些细微的扰动几乎无法被人眼察觉, 但足以影响模型的判断.

此外, 我们还展示了其他典型方案制备的毒化样本视觉效果. 如图 10 所示, 图 10(a) 是 3 种基于 GAN 的毒化样本生成方案, 图 10(b) 则是 3 种基于优化问题的方案, 展示的样本均取自这些方案对应的论文原文 (若效果与扰动参数有关, 则取中等程度的扰动参数). 对比表明, 本文 CPDR 方案在更宽松的假设下, 生成的毒化样本的视觉质量与基于优化问题的方案相当, 具有较强的竞争力.

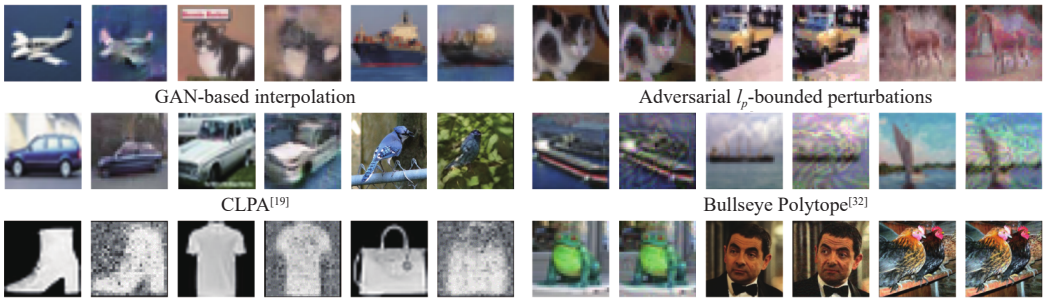
值得注意的是, 为保证视觉隐蔽性, 实验中尽可能促使原始样本向与其深层特征最接近的类别偏移. 具体而言, 计算不同类别间样本特征均值的欧氏距离, 并选择特征空间中最接近的目标类别进行偏移. 由于数据集存在明显差异, 对于 FashionMNIST 数据集, 使用在该数据集上预训练的 EfficientNet-B0 模型提取深层特征; 而对于 Cifar10 数据集, 则采用在该数据集上预训练的 ResNet-50 模型进行特征提取. 图 11 展示了两数据集的类别间特征距离热力图, 直观地展示了两个数据集中各类别在特征空间中分布的近似程度.

最后, 为比较不同方案的执行效率, 本文展示了 4 种不同方案在 Cifar10 数据集上生成 100 个毒化样本的耗

时, 如表 2 所示. 其中 Bullseye Polytope 方案[32]的速度根据其论文给出的数据计算 (生成 10 个样本约 7 min), 其余方案为手动复现. 由表 2 可知, 基于优化问题的方案优势在于无需训练生成模型, 但其在生成阶段耗时较多, 分别在几分钟到几十分钟, 在需要大量制备毒化样本时表现不佳. Bullseye Polytope 方案基于多面体优化数学方法, 计算更加复杂; Adversarial Poison[18] (以下简称 AdvPoison) 则通过投影梯度下降技术显著提高了生成效率. 基于 GAN 的方案和本文方案虽然需要花费数十分钟时间, 但只需训练一次, 之后即可实现毫秒级毒化样本制备. 其中, VagueGAN 在训练时收敛较慢, 但由于生成器结构简单, 生成速度更快; 本文的 CPDR 方案则在训练效率上更有优势, 但生成样本时由于模型结构相对复杂, 略慢于 VagueGAN.

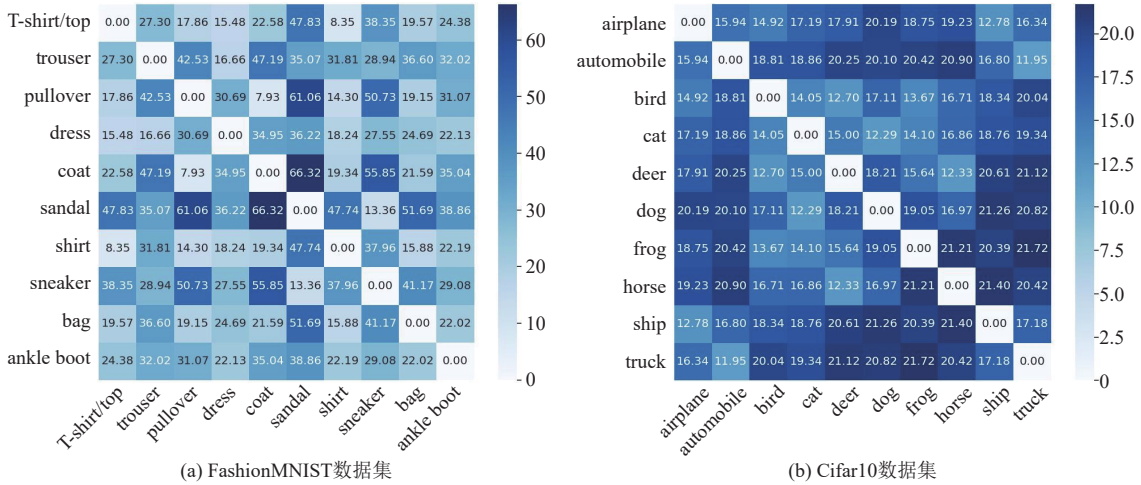
图 8　CPDR 在 FashionMNIST 上生成的毒化样本

图 9　CPDR 在 Cifar10 上生成的毒化样本



(a) 基于GAN的方案 (b) 基于优化问题的方案

图 10 其他典型毒化样本生成方案效果



(a) FashionMNIST数据集 (b) Cifar10数据集

图 11 不同类别特征均值间的欧氏距离

表 2 毒化样本生成效率对比

攻击方案	类型	核心技术	训练耗时 (min)	生成100个毒化样本耗时 (s)
Bullseye Polytope[32]	基于优化问题	多面体优化	—	4200
AdvPoison[18]	基于优化问题	投影梯度下降	—	527
VagueGAN[23]	基于GAN	GAN+自定义损失函数	48	0.23
CPDR	基于数据重建	Attention U-Net+自定义损失函数	23	0.74

4.3 评价指标

CPDR 方案是一个定向投毒攻击方案, 第 4.2 节已直观展示其生成样本视觉上的隐蔽性, 本节采用目标类别准确率 (*Target-Class Accuracy*) 和主任务准确率 (*Main-Task Accuracy*) 作为主要指标来评估其攻击效果.

(1) 目标类别准确率: 用于衡量模型在投毒攻击下对特定目标类别样本的分类准确性. 在定向投毒攻击中, 攻击主要针对一个特定类别, 因此该类别的分类准确率将直接受到影响. 目标类别上的准确率越低, 表明投毒攻击对目标模型的定向破坏效果越明显. 具体而言, 模型在目标类别 y_c 上的准确率计算公式为:

$$Target\text{-}Class\ Accuracy = \frac{\sum_{i=1}^{N} \mathbb{I}_{[y_i=y_c \text{ and } F_M(x_i)=y_c]}(x_i, y_i)}{\sum_{i=1}^{N} \mathbb{I}_{[y_i=y_c]}(x_i, y_i)} \tag{13}$$

其中, N 表示测试集中的样本总数, $F_{\mathrm{M}}(x_i)$ 表示模型对输入样本 x_i 的预测类别, y_i 是 x_i 的真实类别标签, y_i 是目标类别标签. $\mathbb{I}_{[\cdot]}(\cdot)$ 是指示函数, 当方括号内条件为真时取 1, 否则取 0. 因此, 该式分母指测试集中目标类别的样本总数, 分子指测试集中被正确分类的目标类别样本数量.

(2) 主任务准确率: 用于衡量模型在投毒攻击下对所有类别样本的整体分类效果. 定向投毒攻击虽然会影响模型对目标类别的分类能力, 但应尽量减小对其他类别的影响. 主任务准确率越高, 表明投毒攻击的隐蔽性越强, 攻击越难被察觉. 具体而言, 模型主任务准确率的计算公式为:

$$Main\text{-}Task\ Accuracy=\frac{\sum_{i=1}^{N}\mathbb{I}_{[F_{\mathrm{M}}(x_i)=y_i]}(x_i,y_i)}{N} \tag{14}$$

其中, N 表示测试集中的样本总数, $F_{\mathrm{M}}(x_i)$ 表示模型对输入样本 x_i 的预测类别. $\mathbb{I}_{[\cdot]}(\cdot)$ 同样是指示函数, 此处表示对被目标模型正确分类的样本数与全部样本数计算比例.

综上所述, 定向投毒攻击的目标是使模型在目标类别样本上的准确率尽可能低, 同时保持主任务准确率尽可能高, 从而在达成攻击效果的同时增强其隐蔽性.

4.4 集中式学习下的投毒攻击效果

本节在集中式学习场景下评估 CPDR 方案的效果. 具体而言, 将第 4.3 节中展示的毒化样本按照多种比例与正常样本混合, 并在集中式学习场景中发起投毒攻击, 记录中间过程和攻击效果. 本文实验选取 FashionMNIST 数据集中的 T-shirt 类作为目标类别, 诱导模型使其向 shirt 类别偏移. 同时, 将 Cifar10 数据集中的 airplane 类作为目标类别, 使其向 ship 类别偏移. 模型方面, 在 FashionMNIST 数据集上训练 LeNet-5 和 Mini-VGG 模型, 在 Cifar10 数据集上训练 MobileNet-v2 和 ResNet-18 模型. 超参数方面, 实验采用 50 轮的迭代次数、64 的批处理大小 (batch-size)、AdamW 优化器、初始学习率为 0.001 并按余弦退火策略动态调整. 若无特殊说明, 后续实验遵循相同的超参数设置.

图 12 和图 13 分别展示了 CPDR 在 FashionMNIST 和 Cifar10 上的各项基本实验结果, 包括重建模型训练时的损失函数值变化过程, 以及其对不同目标模型的攻击效果. 集中式学习场景不涉及全局模型的更新, 因此仅使用重建损失和投毒效果损失训练重建模型, 并在实验中固定投毒比例为 50%. 由图 12(a) 和图 13(a) 可得, 重建模型在多任务学习的情况下仍保持良好效果, 重建损失和投毒效果损失均呈现下降趋势且最终收敛于一个较小的数值. 值得注意的是, 重建损失明显高于投毒效果损失, 表明重建模型在训练中后期更侧重于优化像素重建任务. 相比于 Cifar10 数据集, 由于 FashionMNIST 数据集中的原始样本均存在大量极端的纯黑色像素, 其重建损失数值相对更高. 进一步地, 由图 12(b)、(c) 和图 13(b)、(c) 可得, 在各种实验设置下, 本方案均可在保证主任务准确率和正常类别 (非目标类别) 分类准确率的情况下, 降低模型对目标类别的分类准确率, 且目标类别和正常类别上的准确率差异明显, 表明了 CPDR 方案的有效性.

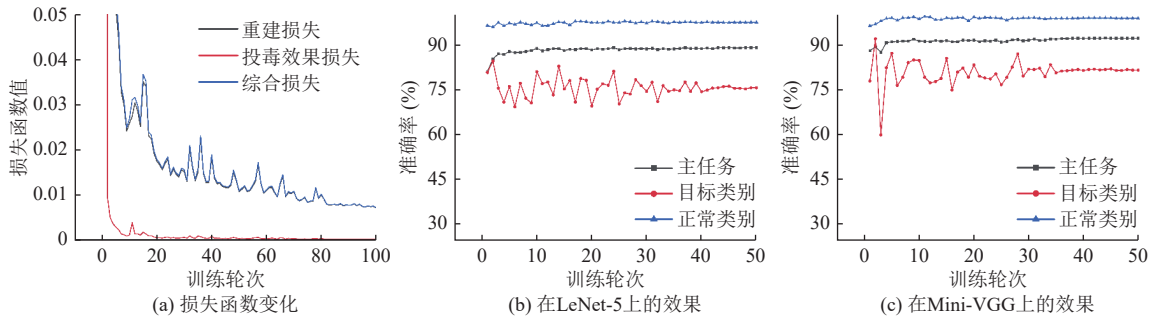
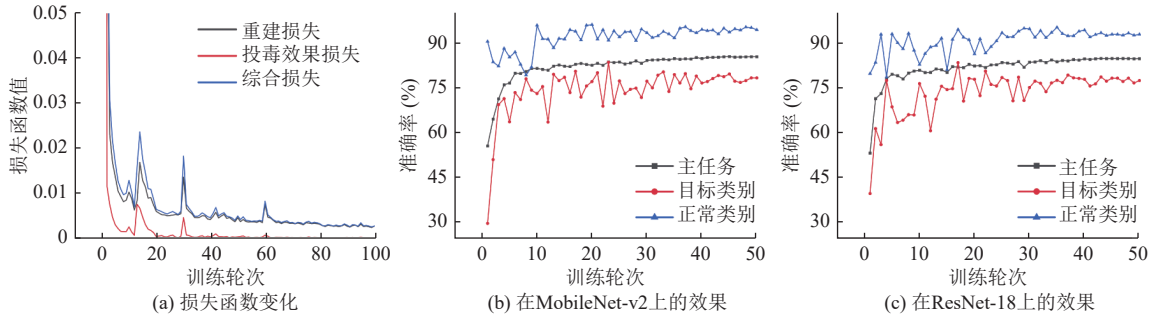


图 12 集中式学习中 FashionMNIST 数据集上的各项实验结果



(a) 损失函数变化 (b) 在MobileNet-v2上的效果 (c) 在ResNet-18上的效果

图 13 集中式学习中 Cifar10 数据集上的各项实验结果

最后, 本节使用 4 种不同数据集和模型的组合, 对比了 CPDR 方案与 VagueGAN[23]和 AdvPoison[18]净标签数据投毒攻击方案在不同投毒率下的效果, 实验结果如表 3 所示, 其中, 加粗数字代表其对应投毒率下的最优攻击效果 (即达成了最低的目标任务准确率). VagueGAN 通过修改条件 GAN 的损失函数, 使其生成具有一定模糊效果的样本以替换正常样本. 本实验取 VagueGAN 中的模糊因子 κ=0.1, 以保证其生成样本尽量与本方案生成样本的视觉质量相近, 而不过于模糊. AdvPoison 则取其官方实现中的设置, 基于 ResNet-18 模型和 PGD 方法生成对抗性样本. 为突出投毒攻击效果, 实验分别评估了 20%、40% 和 60% 投毒率下的结果. 值得说明的是, 此处的投毒率指毒化样本数量在特定类别样本中的比例. 针对全部训练数据而言, 毒化样本所占比例仅为 2%、4% 和 6%, 因此能够保证整体的隐蔽性.

表 3 集中式学习中各类定向投毒攻击下的模型准确率 (%)

数据集模型	任务类型	未投毒	VagueGAN			AdvPoison			CPDR		
			20%	40%	60%	20%	40%	60%	20%	40%	60%
FashionMNIST+LeNet-5	主任务	89.58	89.69	89.84	89.52	89.79	89.33	89.28	90.09	89.82	89.67
	目标任务	83.50	83.30	82.70	77.20	83.30	**79.70**	78.00	**82.60**	79.80	**75.50**
FashionMNIST+Mini-VGG	主任务	92.40	92.59	92.11	91.98	92.51	92.08	91.84	92.48	92.23	91.78
	目标任务	87.70	86.30	84.80	82.20	87.20	83.50	81.80	**85.50**	**82.70**	**79.50**
Cifar10+MobileNet-v2	主任务	86.01	86.36	86.06	85.15	85.68	85.99	85.89	86.48	85.77	85.53
	目标任务	89.20	85.30	83.60	79.60	86.50	83.80	80.10	**83.70**	**80.30**	**76.90**
Cifar10+ResNet-18	主任务	84.81	84.97	84.76	84.15	84.69	84.36	84.23	84.50	84.49	83.88
	目标任务	88.00	86.80	82.80	75.90	86.00	82.90	77.70	**83.20**	**78.50**	**74.80**

对于主任务而言, 上述 3 种方案的总体效果相近. 如表 3 所示, 无论投毒比例如何, 目标模型收敛后的准确率相比于无攻击的情况下波动均在 1% 之内, 表明 3 种方案均具有较强的隐蔽性. 但考虑到毒化样本的质量, AdvPoison 计算出的精确扰动对原始样本影响最小, 肉眼几乎不可判别, 隐蔽性最强; CPDR 存在微小的扰动痕迹, 隐蔽性其次; 相比之下, VagueGAN 存在明显的模糊效果, 隐蔽性弱于 AdvPoison 和 CPDR.

对于目标任务而言, 本文的 CPDR 方案在绝大多数设置下都领先 VagueGAN 和 AdvPoison 方案. 如果将一个数据集和目标模型的组合作为一个任务, 不同投毒比例下主任务与目标任务准确率的平均差值作为该任务下攻击效果的度量指标, 则 CPDR 方案在表 3 中 4 个任务下的平均效果分别为 10.56%、9.60%、5.63% 和 5.76%, 分别领先 VagueGAN 方案 1.94%、1.81%、2.61%、2.97%, 并领先 AdvPoison 方案 1.43%、1.62%、3.08%、3.53%. 该结果表明, 在定向投毒攻击方面, 精心设计的特征偏移在一定程度上比少量不可控的模糊效果和微小的对抗性扰动更加有效.

4.5 联邦学习下的投毒攻击效果

本节在联邦学习场景下评估 CPDR 方案的效果, 并仍与净标签投毒攻击方案 VagueGAN 和 AdvPoison 进行对比. 本节所有实验的批处理大小 (batch size) 均设置为 64, 学习率均设置为 0.001, 使用 AdamW 优化器进行训练,

这些设置与第 4.4 节相同. 在联邦学习中, 设置 10 个参与方, 按照独立同分布设置划分训练数据集, 每个参与方本地训练迭代次数为 2 轮. 为确保模型收敛, FashionMNIST 数据集上的实验全局训练迭代次数为 10 轮, Cifar10 数据集上的实验全局训练迭代次数则为 30 轮. 考虑到联邦学习中的恶意参与方不超过一半, 因此本文在 0–50% 的恶意客户端比例下进行投毒攻击实验. 联邦学习中每个参与方分得的样本较少, 目标类别的样本则更少, 因此为保证攻击强度, 此处不再设定投毒比例, 而是令恶意参与方将其持有的目标类别样本全部替换为毒化样本, 该设置与文献 [12] 保持一致. 最终, 在上述不同的数据集和模型设置下, 统计并分析当恶意客户端比例变化时, 模型在主任务上的准确率和投毒目标类别上的准确率.

实验结果如图 14 所示, 4 幅子图分别展示了不同数据集和模型组合下的投毒攻击效果对比. 其中, 虚线表示模型在主任务上的准确率, 实线表示模型在投毒目标类别上的准确率. 实验结果表明, 随着恶意客户端比例的增加, 各类攻击方案的有效性整体呈现提升趋势, 且对主任务准确率的影响十分有限. 进一步分析不同方案的攻击效果, 可得出以下结论. 首先, 与 VagueGAN 相比, CPDR 在 4 个实验任务中分别领先其 4.16%、4.32%、3.69%、5.58%, 且随着恶意客户端比例增加, 二者的攻击效果差距呈现扩大趋势. 究其原因, 主要与 GAN 的固有特性有关. VagueGAN 通过引入模糊因子 κ 控制生成样本的扰动程度, 但由于 GAN 本身训练不稳定, 过大的模糊因子 κ 会导致生成器难以收敛, 而过小的模糊因子 κ 无法提供足够的攻击信息量, 即引导深层特征定向偏移的噪声量不足. 相比之下, CPDR 能够稳定生成引导深层特征定向偏移的噪声, 攻击稳健性更高. 与 AdvPoison 相比, CPDR 在 4 个实验任务中分别领先其 0.92%、3.86%、4.27%、7.34%. 虽然 AdvPoison 通过投影梯度下降算法实现了对单样本扰动的精确控制, 具有更优的视觉隐蔽性, 但 CPDR 通过重建模型实现的特征偏移在联邦学习聚合过程中展现出更强的投毒攻击效果. 综上所述, CPDR 方案能够在具有一定隐蔽性的情况下达到较优的净标签定向投毒攻击效果, 在隐蔽性和投毒效果之间取得了良好的平衡.

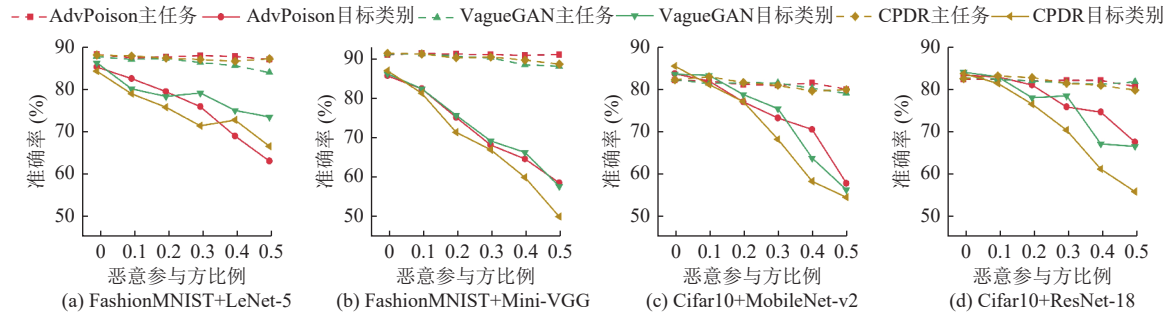


图 14 联邦学习中各类定向投毒攻击下的模型准确率

5 总 结

本文提出一种基于数据重建的深度学习净标签数据投毒攻击方案 CPDR. 该方案将 Attention U-Net 作为图像重建模型, 并重构其损失函数, 使其能够在保持输入样本视觉特征稳定性的前提下, 促使重建样本的深层特征发生偏移, 从而生成高隐蔽性的毒化样本以实施投毒攻击. 实验结果表明, 该方案生成的毒化样本在视觉上与原始样本非常相似, 能够在几乎不影响深度学习主任务表现的情况下有效降低模型在目标类别上的准确率. 在未来, 相关研究可进一步拓展至多模态数据场景, 并在多种防御机制下综合评估攻击效果, 以实现隐蔽性与效率之间的更优平衡.

References

[1] Badue C, Guidolini R, Carneiro RV, Azevedo P, Cardoso VB, Forechi A, Jesus L, Berriel R, Paixão TM, Mutz F, de Paula Veronese L, Oliveira-Santos T, De Souza AF. Self-driving cars: A survey. Expert Systems with Applications, 2021, 165: 113816. [doi: 10.1016/j.eswa.2020.113816]

[2] Ozbayoglu AM, Gudelek MU, Sezer OB. Deep learning for financial applications: A survey. Applied Soft Computing, 2020, 93: 106384. [doi: 10.1016/j.asoc.2020.106384]

[3] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. Medical Image Analysis, 2017, 42: 60–88. [doi: 10.1016/j.media.2017.07.005]

[4] Chen HS, Liu XR, Yin DW, Tang JL. A survey on dialogue systems: Recent advances and new frontiers. ACM SIGKDD Explorations Newsletter, 2017, 19(2): 25–35. [doi: 10.1145/3166054.3166058]

[5] Tramèr F, Zhang F, Juels A, Reiter MK, Ristenpart T. Stealing machine learning models via prediction APIs. In: Proc. of the 25th USENIX Conf. on Security Symp. Austin: USENIX Association, 2016. 601–618.

[6] Shokri R, Stronati M, Song CZ, Shmatikov V. Membership inference attacks against machine learning models. In: Proc. of the 2017 IEEE Symp. on Security and Privacy. San Jose: IEEE, 2017. 3–18. [doi: 10.1109/SP.2017.41]

[7] Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: Proc. of the 3rd Int'l Conf. on Learning Representations. San Diego: ICLR, 2015.

[8] Biggio B, Nelson B, Laskov P. Poisoning attacks against support vector machines. In: Proc. of the 29th Int'l Conf. on Machine Learning. New York: ICML, 2012. 1807–1814.

[9] Gao Y, Chen XF, Zhang YY, Wang W, Deng HH, Duan P, Chen PX. A survey of attack and defense techniques for federated learning systems. Chinese Journal of Computers, 2023, 46(9): 1781–1805 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2023.01781]

[10] Liu YQ, Ma SQ, Aafer Y, Lee WC, Zhai J, Wang WH, Zhang XY. Trojaning attack on neural networks. In: Proc. of the 25th Annual Network and Distributed System Security Symp. San Diego: The Internet Society, 2018. 301–315.

[11] Bagdasaryan E, Veit A, Hua YQ, Estrin D, Shmatikov V. How to backdoor federated learning. In: Proc. of the 23th Int'l Conf. on Artificial Intelligence and Statistics. Palermo: PMLR, 2020. 2938–2948.

[12] Tolpegin V, Truex S, Gursoy ME, Liu L. Data poisoning attacks against federated learning systems. In: Proc. of the 25th European Symp. on Research in Computer Security. Guildford: Springer, 2020. 480–501. [doi: 10.1007/978-3-030-58951-6_24]

[13] Xiao X, Tang Z, Li CY, Xiao B, Li KL. SCA: Sybil-based collusion attacks of IIoT data poisoning in federated learning. IEEE Trans. on Industrial Informatics, 2023, 19(3): 2608–2618. [doi: 10.1109/TII.2022.3172310]

[14] Gupta P, Yadav K, Gupta BB, Alazab M, Gadekallu TR. A novel data poisoning attack in federated learning based on inverted loss function. Computers & Security, 2023, 130: 103270. [doi: 10.1016/j.cose.2023.103270]

[15] Wang RJ, Wang JB, Zhang FL, Li JW, Li ZP, Chen T. Feature map poisoning attack and dual defense mechanism for federated prototype learning. Ruan Jian Xue Bao/Journal of Software, 2025, 36(3): 1355–1374 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/7183.htm [doi: 10.13328/j.cnki.jos.007183]

[16] Steinhardt J, Koh PW, Liang P. Certified defenses for data poisoning attacks. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 3520–3532.

[17] Shafahi A, Huang WR, Najibi M, Suciu O, Studer C, Dumitras T, Goldstein T. Poison frogs! Targeted clean-label poisoning attacks on neural networks. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018. 6106–6116.

[18] 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. Curran Associates Inc., 2021. 2321.

[19] Zhao BY, Lao YJ. CLPA: Clean-label poisoning availability attacks using generative adversarial nets. In: Proc. of the 36th AAAI Conf. on Artificial Intelligence. AAAI, 2022. 9162–9170. [doi: 10.1609/aaai.v36i8.20902]

[20] Koh PW, Steinhardt J, Liang P. Stronger data poisoning attacks break data sanitization defenses. Machine Learning, 2022, 111(1): 1–47. [doi: 10.1007/s10994-021-06119-y]

[21] Zeng Y, Pan MZ, Just HA, Lyu LJ, Qiu MK, Jia RX. Narcissus: A practical clean-label backdoor attack with limited information. In: Proc. of the 2023 ACM SIGSAC Conf. on Computer and Communications Security. Copenhagen: ACM, 2023. 771–785. [doi: 10.1145/3576915.3616617]

[22] Yang J, Zheng J, Baker T, Tang S, Tan YA, Zhang QX. Clean-label poisoning attacks on federated learning for IoT. Expert Systems, 2023, 40(5): e13161. [doi: 10.1111/exsy.13161]

[23] Sun W, Gao B, Xiong K, Lu Y, Wang YW. VagueGAN: A GAN-based data poisoning attack against federated learning systems. In: Proc. of the 20th Annual IEEE Int'l Conf. on Sensing, Communication, and Networking. Madrid: IEEE, 2023. 321–329. [doi: 10.1109/SECON58729.2023.10287523]

[24] Alharbi S, Guo YF, Yu W. Collusive backdoor attacks in federated learning frameworks for IoT systems. IEEE Internet of Things

Journal, 2024, 11(11): 19694–19707. [doi: 10.1109/JIOT.2024.3368754]

[25] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: Proc. of the 18th Int'l Conf. on Medical Image Computing and Computer-assisted Intervention. Munich: Springer, 2015. 234–241. [doi: 10.1007/978-3-319-24574-4_28]

[26] Oktay O, Schlemper J, Le Folgoc L, Lee M, Heinrich M, Misawa K, Mori K, McDonagh S, Hammerla NY, Kainz B, Glocker B, Rueckert D. Attention U-Net: Learning where to look for the pancreas. In: Proc. of the 1st Int'l Conf. on Medical Imaging with Deep Learning. Amsterdam: OpenReview.net, 2018.

[27] Chowdhury A, Jiang MC, Chaudhuri S, Jermaine C. Few-shot image classification: Just use a library of pre-trained feature extractors and a simple classifier. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 9425–9434. [doi: 10.1109/ICCV48922.2021.00931]

[28] Papyan V, Han XY, Donoho DL. Prevalence of neural collapse during the terminal phase of deep learning training. Proc. of the National Academy of Sciences of the United States of America, 2020, 117(40): 24652–24663. [doi: 10.1073/pnas.2015509117]

[29] Rangamani A, Lindegaard M, Galanti T, Poggio T. Feature learning in deep classifiers through intermediate neural collapse. In: Proc. of the 40th Int'l Conf. on Machine Learning. Honolulu: JMLR.org, 2023. 1194.

[30] Schroff F, Kalenichenko D, Philbin J. FaceNet: A unified embedding for face recognition and clustering. In: Proc. of the 2015 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 815–823. [doi: 10.1109/CVPR.2015.7298682]

[31] Sharma S, Xian YQ, Yu N, Singh A. Learning prototype classifiers for long-tailed recognition. In: Proc. of the 32nd Int'l Joint Conf. on Artificial Intelligence. Macao: ijcai.org, 2023. 151. [doi: 10.24963/ijcai.2023/151]

[32] Aghakhani H, Meng DY, Wang YX, Kruegel C, Vigna G. Bullseye polytope: A scalable clean-label poisoning attack with improved transferability. In: Proc. of the 2021 IEEE European Symp. on Security and Privacy. Vienna: IEEE, 2021. 159–178. [doi: 10.1109/EuroSP51992.2021.00021]

附中文参考文献

[9] 高莹, 陈晓峰, 张一余, 王玮, 邓煌昊, 段培, 陈培炫. 联邦学习系统攻击与防御技术研究综述. 计算机学报, 2023, 46(9): 1781–1805. [doi: 10.11897/SP.J.1016.2023.01781]

[15] 王瑞锦, 王金波, 张凤荔, 李经纬, 李增鹏, 陈厅. 联邦原型学习的特征图中毒攻击和双重防御机制. 软件学报, 2025, 36(3): 1355–1374. http://www.jos.org.cn/1000-9825/7183.htm [doi: 10.13328/j.cnki.jos.007183]

作者简介

张文博, 硕士, 主要研究领域为深度学习安全.

李雄, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为应用密码学, 数据安全, 隐私计算.

张文琪, 博士生, 主要研究领域为隐私计算.

谢勇, 博士, 教授, CCF 高级会员, 主要研究领域为轻量化人工智能, 物联网应用技术.

陈厅, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为区块链安全, 软件安全, 软件工程.

禹继国, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为网络与数据安全, 区块链, 隐私计算.

张小松, 博士, 教授, 博士生导师, CCF 会士, 主要研究领域为网络安全, 卫星互联网安全, 区块链.