

基于扩散模型的机器学习后门攻击防御方法

王孙平¹, 张爱清¹, 叶新荣¹, 王勇²

(1. 安徽师范大学物理与电子信息学院, 安徽 芜湖 241002;

2. 皖南医学院医学信息学院, 安徽 芜湖 241002)

摘要: 后门攻击是机器学习模型在训练阶段面临的主要安全威胁之一。尽管现有针对后门攻击的防御研究已取得显著进展, 但这些方法往往会导致模型在干净测试集上的准确率出现明显下降。为解决这一问题, 提出了一种基于扩散模型的机器学习后门攻击防御方法 (defending against backdoor attacks with diffusion model, DBADM)。该方法的核心思想是在模型训练前, 利用扩散模型对含有后门触发器的中毒样本进行预处理, 通过改变样本中隐藏的触发器特征, 从而有效抵御后门攻击。研究人员在MNIST、CIFAR-10、Tiny ImageNet和LFW4个基准数据集上进行了系统的攻防对比实验。实验结果表明, 所提出的DBADM方法不仅能够有效防御各类后门攻击, 还能保持模型在干净数据集上的高精度性能。

关键词: 机器学习; 后门攻击; 触发器; 扩散模型

中图分类号: TP393

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2025059

method for machine learning backdoor attacks based on defense method based on diffusion models

WANG Sunping¹, ZHANG Aiqing¹, YE Xinrong¹, WANG Yong²

1. School of Physics and Electronic Information, Anhui Normal University, Wuhu 241002, China

2. School of Medical Information, Wannan Medical College, Wuhu 241002, China

Abstract: Backdoor attacks were recognized as one of the primary security threats faced by machine learning models during the training phase. Although significant progress had been achieved in existing defense methods against backdoor attacks, these approaches were often found to result in a substantial decline in model accuracy on clean test sets. To address this issue, a method named defending against backdoor attacks with diffusion model (DBADM) was proposed. The core idea of this method was to preprocess poisoned samples containing backdoor triggers using a diffusion model before model training. By altering the hidden trigger features in the samples, back-

收稿日期: 2025-03-23; 修回日期: 2025-06-26

通信作者: 张爱清, aqzhang2006@163.com

基金项目: 国家自然科学基金项目 (62072005); 安徽省自然科学基金优秀青年科学基金项目 (2108085Y22)

Foundation Items: The National Natural Science Foundation of China (62072005), Anhui Provincial Natural Science Foundation for Outstanding Young Scientists Project (2108085Y22)

引用格式: 王孙平, 张爱清, 叶新荣, 等. 基于扩散模型的机器学习后门攻击防御方法[J]. 网络与信息安全学报, 2025, 11(5): 137-148.

Citation Format: WANG S P, ZHANG A Q, YE X R, et al. Defense method for machine learning backdoor attacks based on diffusion model[J]. Chinese Journal of Network and Information Security, 2025, 11(5): 137-148.

door attacks were effectively mitigated. Systematic offensive and defensive comparison experiments were conducted on four benchmark datasets: MNIST, CIFAR-10, Tiny ImageNet, and LFW. The experimental results demonstrate that the DBADM method not only successfully defends against various backdoor attacks but also maintains the model's high accuracy performance on the clean dataset.

Keywords: machine learning, backdoor attack, trigger, diffusion model

0 引言

随着人工智能技术的飞速发展,深度神经网络(deep neural networks, DNN)作为机器学习(machine learning, ML)的核心技术,已在图像识别、自然语言处理、自动驾驶、医疗诊断等领域展现出强大的应用潜力。这类网络通过模拟人脑神经元的结构与功能,能够从海量数据中自动提取特征并进行预测,具有强大的数据表达能力。然而,训练高性能的深度学习模型通常需要消耗大量的数据和计算资源,这让许多企业和研究机构面临资源不足的困境。为应对这一挑战,模型训练任务外包和使用第三方数据训练模型逐渐成为行业常态,这种模式虽然降低了机器学习模型开发的门槛和成本,让众多缺乏充足资源的小型企业得以运用先进的机器学习技术,但也随之引入了新的安全隐患,尤其在企业遭遇恶意攻击时,相关风险更加凸显。

在机器学习模型训练过程中,攻击者可能通过篡改训练数据或植入后门等方式干扰模型的学习过程。后门攻击是一种典型的威胁方式,攻击者在训练数据中巧妙地加入带有特定触发器的样本,这些样本在训练阶段通常不会影响模型的正常学习,但当模型部署到实际应用环境后,一旦接收到包含相同触发器的输入,就会按照攻击者预设的意图输出错误分类结果。这种攻击的危害不只局限于模型本身的安全失效,还可能引发连锁性的社会安全风险。以金融领域的实际场景为例,攻击者通过篡改欺诈检测模型的训练数据,使模型在训练过程中形成错误的判断逻辑,将本应识别的欺诈行为误判为正常交易,最终造成重大的经济损失。

当前,机器学习模型已在诸多关键领域得到深度应用,如自动驾驶、智慧医疗、金融欺诈检测及人脸识别等^[1]。以自动驾驶系统为例,其核心依赖深度学习模型完成交通标志识别、行人与障碍物检测等任务,并基于识别结果进行驾驶决

策。一旦这类模型遭遇后门攻击,可能会引发交通事故,甚至造成人员伤亡。类似地,人脸识别技术被广泛应用于安全验证与身份认证场景,若模型被植入后门,攻击者可能通过预设条件绕过验证机制,进而引发身份盗用、非法访问等安全事件。因此,研究机器学习模型的后门攻击防御方法、保障模型安全,已成为人工智能领域一个至关重要的研究方向。

近年来,针对深度学习的后门攻击防御研究取得了一定进展^[2]。然而,现有防御方法普遍面临检测精度与模型性能的权衡问题。主流防御技术通常通过检测并清除训练数据中的潜在后门触发器,防止恶意模型的生成。常见的检测手段包括基于样本特征的异常检测^[3]、基于梯度变化的监控^[4]等。这些方法通过分析模型的训练过程来识别模型可能受攻击的迹象。但这类防御方法往往存在“过度清除”的风险,即可能将正常数据特征误判为后门攻击相关的异常信号,进而对正常样本进行剔除或弱化处理,最终导致模型的测试精度下降。尤其当后门攻击的触发器设计得极为微小或隐蔽时,现有防御方法的检测效果往往不尽如人意^[5]。文献[6]的实验表明,在理想条件下,即便选取NAD(neural attention distillation)^[6]和Fine-pruning^[7]两种方法的最佳实验结果,其模型精度相比无防御措施的基准模型,仍分别下降了4.48%和4.51%。因此,如何在有效防御后门攻击的同时最小化模型性能损失,成为当前后门防御领域研究的核心挑战。

针对这一问题,一种新兴的防御思路是在数据处理环节引入扩散模型。作为一种基于概率的生成模型,扩散模型通过“前向加噪”与“逆向去噪”两个核心过程实现数据的特征变换。在前向过程中,输入数据会被逐步添加噪声,最终转化为纯随机噪声;在逆向过程中,模型则通过逐步去噪操作,将随机噪声还原为清晰的数据。Ho等^[8]提出的去噪扩散概率模型(denoising dif-

fusion probabilistic model, DDPM) 在去噪过程中能够逐步恢复图像的噪声特征, 而面对后门攻击时, 可以利用这一特性抑制触发器的影响。具体而言, 后门攻击中的触发器可被视为图像中的一种“异常噪声”, 在扩散模型的逆向去噪过程中, 触发器的位置会因噪声消除而发生偏移, 部分图像中的触发器甚至会变得模糊(或无法维持原有形态)。通过这种方式, 触发器与模型之间的关联被切断, 后门攻击的影响得以有效消除。此外, Zhou等^[9]提出利用扩散模型清除中毒图像中潜藏的触发器特征, 实现对中毒数据集的净化。该方案的核心优势在于, 能够在剔除触发器的同时保留图像原有的正常特征, 避免对有效数据造成破坏。不过, 该方案需要对超过65%的图像样本进行抽样处理才能达到预期效果。这一高比例抽样要求不仅会增加数据处理的计算开销, 还可能在样本总量有限的场景中导致有效数据过度消耗, 从而制约了其在实际场景中的推广应用。

综上所述, 随着机器学习技术在各领域的广泛应用, 开发一种既能有效抵御后门攻击, 又能避免损害模型原有性能的防御方法已成为当务之急。为此, 本文提出一种基于扩散模型的后门攻击防御方法。

本文的主要贡献如下。

1) 提出基于扩散模型的后门攻击防御(defending against backdoor attacks with diffusion model, DBADM)方法。该方法针对被攻击目标进行处理, 使被攻击者上触发器的位置和形状发生变化, 从而达到抵御后门攻击的目的。

2) 在MNIST、CIFAR-10、Tiny ImageNet及LFW 4个公开数据集上对DBADM方法进行了系统性测试。实验结果表明, 该方法不仅能够抵御后门攻击, 还可以保证模型在干净数据集上保持较高的测试精度。

1 相关工作

1.1 后门攻击

后门攻击, 又称木马攻击, 是一种通过在机器学习模型中植入特定触发机制, 使其在面对含有预设触发器的输入时产生目标输出的安全威胁。从攻击对象的角度划分, 机器学习领域的后门攻击可分为数据投毒攻击与模型投毒攻击两

类。本文研究的BadNets^[10]属于典型的数据投毒攻击。它由Gu等^[10]提出, 是机器学习领域最早被提出的后门攻击。其攻击流程如下: 首先从干净训练集中选取部分样本, 为其添加给定的触发器后, 用目标标签替换样本原始标签以生成有毒样本; 随后将有毒样本混入干净数据集, 形成混合训练集用于模型训练, 最终使模型在含触发器的输入下输出攻击者预设的错误结果。

在不可见后门攻击研究领域, 学者提出了多种创新性方法。Ma等^[11]设计了一种动态不可见触发器算法, 其核心思路是先通过算法定位神经网络的决策边界, 再将触发器与图像内容深度融合。这种设计既保证了触发器的隐蔽性(难以被人工或自动化检测发现), 又能在低投毒比率下实现高攻击成功率。实验数据显示, 即便在训练集中未注入任何有毒样本(投毒率为0%)的极端条件下, 该方法仍能成功实施后门攻击。Zhong等^[12]将触发器视为一种遵循多项式分布的特殊噪声, 并利用U-net为每个正常输入生成适配的多项式分布参数。这一设计使触发器在人眼视觉层面与正常图像特征无明显差异, 同时规避了统计检测工具的识别。进一步地, 他们通过在特征表示空间中设计正则化项, 使含触发器的恶意输入与正常输入的特征表示高度纠缠, 增强了对模型诊断防御的鲁棒性。Gao等^[13]提出一种稀疏且不可见后门攻击方法, 采用优化算法生成兼具稀疏性与视觉隐蔽性的触发器。这种设计在保证触发器难以被察觉的同时, 有效提高了攻击的成功率。

尽管不可见后门攻击能使中毒图像在视觉上与正常图像无明显差异, 但由于其标签与正常样本的标签存在差异, 通过分析训练样本中图像一标签的对应关系仍可能被检测到。为进一步提升攻击的隐蔽性, 研究者提出了干净标签攻击这一新方向。Zhao等^[14]首次设计了一种鲁棒的基于少样本学习的语句改写模型。该模型能够生成标签正确的有毒样本, 并以文本样式作为触发器。该触发器相比传统的触发器更加抽象, 更难被检测工具识别, 显著提升了攻击的隐蔽性。Gao等^[15]深入分析了干净标签攻击的核心难点, 即中毒样本需包含与目标标签相关的鲁棒特征, 但

不同样本的鲁棒特征对模型而言学习难度差异显著。因此,他们提出使用损失值、梯度范数和遗忘事件等指标筛选“难学”样本作为攻击目标的策略。实验结果显示,针对这类样本植入后门,能够有效提升攻击成功率。Zeng 等^[16]针对现有干净标签攻击方法需要全局训练集知识的局限性,提出了一种新型后门攻击方法,该方法能够在极低的投毒比条件下无须获取完整的训练集信息便可实现有效的后门攻击。

上述研究表明,后门攻击发展迅速,已衍生出多种形态,从早期可见触发器的简单设计,到不可见触发器的隐蔽化改进,再到干净标签攻击的无标签异常特征设计,攻击手段的隐蔽性、有效性和适用性持续提升。这些发展趋势不仅对机器学习模型的安全构成了更严峻的威胁,也对现有防御技术提出了更高要求。因此,设计一种能覆盖多类攻击场景的通用防御方法,已成为保障机器学习应用安全的迫切需求。

1.2 后门防御

后门攻击的技术手段日益多样化,与之相对应的防御策略也在不断丰富。目前,后门攻击的主流防御方法可分为后门检测和触发器擦除^[6]两大类。这两类方法分别从识别攻击和消除攻击影响两个维度构建防御体系,具体技术路径如下。

(1) 基于后门检测的防御方法

此类方法的核心目标是识别潜在的后门攻击或过滤输入数据中的恶意样本。Gao 等^[17]针对计算机视觉任务提出了一种后门识别方法,但该方法难以推广到其他领域。为此,他们进一步优化设计了一种跨领域防御方法。该方法无须依赖机器学习专业知识,也不需要昂贵的计算资源支持,通过通用特征分析即可实现对视觉、文本、音频等的后门攻击检测,显著提升了实用性。Li 等^[18]利用后门攻击的非转移性特性,提出了一种基于特征相似性的检测方法。具体而言,该方法先使用特征提取器获取输入数据的特征向量,再从中毒模型预测的类别标签中随机选取一组样本,通过对比输入数据与样本组在特征提取器潜在空间中的相似度,判断输入是否为干净数据。

(2) 基于触发器擦除的防御方法

此类方法聚焦于攻击影响消除,通过直接净

化被植入后门的模型或数据,消除触发器与错误输出之间的关联,同时尽可能保留模型对正常数据的识别能力。Chen 等^[19]通过机理分析发现,后门的生效本质是代表后门逻辑的神经元簇与触发器特征建立了强连接。基于这一发现,他们设计了 LinkBreaker 方法,通过识别并切断与触发器相关的神经元连接,从而抵御后门攻击。Zhang 等^[20]采用模型解耦策略,将模型分解为特征提取器和分类器两部分,并利用解耦对比学习破坏触发器特征与目标标签之间的相关性。在此基础上,他们提出了一种自适应分类器聚合策略,根据数据特征动态调整分类器权重系数,进一步降低触发器对预测结果的干扰。

随着后门攻击在机器学习领域引发的关注持续升温,相关防御技术也随之取得了显著进步,目前已形成多种防御策略,例如,从被污染的训练数据集中精准识别并删除有毒样本,或在模型训练过程中实时过滤可能携带攻击特征的数据。尽管这些方法在抵御后门攻击方面展现出良好效果,但普遍存在一个关键问题:在提升模型安全性的同时,往往会导致模型在干净数据集上的测试精度出现不同程度的下降。这种防御效果与模型性能之间的权衡关系,成为当前后门防御研究面临的核心挑战。因此,本文的研究重点在于探索如何在确保有效防御后门攻击的同时,保持模型测试精度。

2 威胁模型与防御方法

2.1 威胁模型

本文采用与传统的后门攻击方法 BadNets^[10]一致的攻击模型,其攻击过程可以划分为以下几个阶段。

1) 中毒样本生成。攻击者首先从干净样本集中选取目标投毒样本,在图像的边角等非关键区域添加特定的触发器。这些触发器通常设计为微小且视觉隐蔽的图案,既能规避人工识别,又不会显著改变图像的核心特征。随后,攻击者将这些含触发器图像的原始标签替换为预设的目标标签,最终生成“触发器+错误标签”绑定的中毒样本。

2) 后门模型训练。攻击者将中毒样本混入干净样本集,形成被污染的训练数据集用于模型训

练。在模型训练过程中，由于中毒样本反复输入，模型会在学习正常数据特征的同时，被动“记住”触发器与目标标签之间的关联。这种关联本质上是人为植入的错误逻辑（触发器作为无关外部标记，本不应影响模型决策），但模型会将其视为“有效特征”存储于参数中，最终完成后门模型。

3) 后门触发与生效。经过训练的后门模型在处理干净输入（无触发器的正常样本）时，能够基于正常特征做出准确预测，其表现与普通模型无异；然而，当输入包含预设触发器时，模型会立即激活训练阶段学到的错误关联模式。此时，模型将完全忽略输入数据的真实特征，强制输出攻击者预设的目标标签。这一机制的核心在于，触发器会特异性激活模型中通过后门训练形成的后门神经元^[7]。这类神经元不参与正常数据的特征处理，仅在检测到触发器时被激活，从而直接干扰模型的正常决策逻辑。正因如此，后门模型虽在日常使用中表现正常，但在触发条件满足时会突然失效，给系统安全带来重大隐患。

2.2 防御方法

DDPM被广泛应用于图像生成领域，其核心机制由前向扩散过程与逆向去噪过程两部分构成。

在前向扩散过程中，逐步向原始数据中添加高斯噪声，直到数据完全变为噪声。具体而言，假设训练集中真实图像的数据分布为 $q(x_0)$ ，从中随机抽取一幅原始图像，表示为 $x_0 \sim q(x_0)$ 。前向过程 $q(x_t|x_{t-1})$ 是指在前向的每一个扩散步中，系统按照预设的噪声调度方案对图像 x_{t-1} 添加高斯噪声，获得噪声增强的图像 x_t 。值得注意的是，整个前向过程具有马尔可夫链特性，即当前

状态仅依赖于前一步状态。

逆向去噪过程是前向扩散的逆操作，即逐步去噪的过程。在时刻 T 从标准高斯分布中随机采样一个二维高斯噪声，然后逐步去噪，最终生成图像 x_0 与真实图像的分布一致。

扩散模型的核心能力在于精准去噪，这为后门攻击防御提供了新思路。本文将后门攻击图像中的触发器视为一种异常噪声，利用扩散模型的去噪机制对其进行处理，从而实现抵御后门攻击的目标。具体而言，触发器作为嵌入图像的特殊噪声，其本质是通过干扰模型特征提取过程诱导错误决策；而扩散模型可以有效减弱或消除触发器的影响，提升模型的安全性。由于本文的核心研究内容是探究扩散模型能否通过改变触发器的位置与形状来抵御后门攻击，因此本文所提方法的模型以“已知中毒样本”为前提条件。

DBADM方法的模型结构如图1所示。在模型正式训练前，先使用扩散模型对训练集中的中毒样本进行训练，再将扩散模型生成的样本加入训练集，形成净化后的训练数据。图2展示了原始图像样本及经扩散模型处理后的效果。通过对处理后样本的统计与观察，发现扩散模型对中毒样本的影响主要表现为3种情况，即图片保持原样、触发器的位置发生变化和触发器变得模糊（或无法保持原有形状）。其中，“触发器的位置发生变化”占比最高（为50%~60%）。使用上述被扩散模型处理后的数据集进行模型训练，可以起到抵御后门攻击的作用。实验表明，该方案对可见后门攻击具有显著防御效果，但对不可见后门攻击的防御效果有限。原因有二：一是不可见后门攻击的触发器非常隐蔽，扩散模型的全图去噪难以精准定位并削弱其影响；二是对整张图

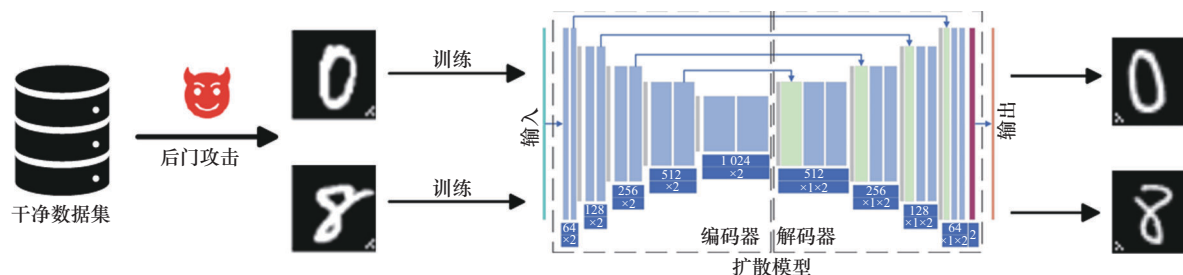


图1 DBADM方法的模型结构

Figure 1 Model structure of the DBADM method

像的无差别处理可能同时弱化正常特征，反而影响模型性能。

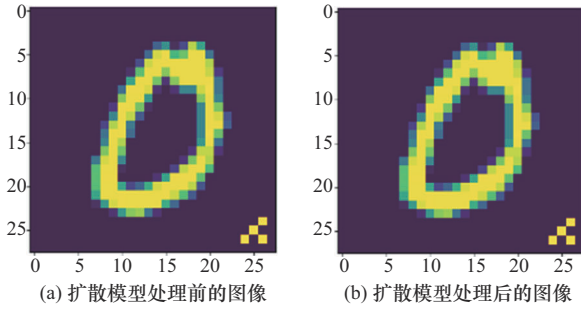


图2 扩散模型处理前后的图像对比

Figure 2 Comparison of images before and after diffusion model processing

为解决上述问题，本文考虑引入RISE (randomized input sampling for explanation) 方法进行优化。RISE的核心原理是通过随机掩码对输入图像进行扰动，观察模型输出的变化，生成反映像素重要性的热力图。触发器像素因能显著影响模型输出，会在热力图中呈现高权重特征。基于此，可先通过RISE对中毒图像中的触发器进行定位，再用扩散模型对该区域进行针对性处理，从而在保留正常特征的同时精准消除不可见触发器的影响，提升对隐蔽性攻击的防御能力。

3 实验与结果分析

本节首先介绍实验环境及参数设置，然后分析DBADM、Fine-pruning、NAD^[6]和DataElixir^[9]对比实验结果。所有的实验在NVIDIA RTX 4090 GPU上完成。

3.1 实验设置

为全面验证DBADM方法的防御效果，实验选取MNIST、CIFAR-10、Tiny ImageNet和LFW 4个主流图像数据集作为测试载体，并在数据集中分别注入可见后门攻击BadNets^[10]和不可见后门攻击Blended^[21]。在实验过程中，将所有数据集中的中毒图像统一调整为32×32像素，并对这些中毒图像进行预处理，将原始PNG格式图像转换为NumPy数组，然后进行归一化处理，最后将NumPy数组转换为PyTorch张量，作为模型输入数据格式。此外，将扩散模型的步数 T 设置为1 000（可运用DDIM算法加速采样过程），批

次大小设置为64（常规设置为128，但此值设置得越大，对显存的要求越高），学习率设置为0.000 2，损失函数选用L2（即均方误差），训练步数为70万。

(1) 扩散模型选择

实验对比了3种主流扩散模型的适用性，分别为Ho等^[8]提出的DDPM、Nichol等^[22]提出的Improved DDPM和Dhariwal等^[23]提出的ADM。结果如表1所示。

表1 DBADM下不同扩散模型的攻击成功率和生成的图像质量

Table 2 ASR and image generation quality of different diffusion models under DBADM

扩散模型	攻击成功率	图像的质量
DDPM	0.10	一般
Improved DDPM	0.08	优秀
ADM	0.075	0.075

1) DDPM：在低分辨率（如32×32、64×64）图像生成任务中表现不佳，且抵御后门攻击的效果最差。

2) ADM：在防御效果上与Improved DDPM不相上下，但ADM为提升生成的图像质量采用了“增加网络深度+增加注意力头数”的设计，这导致计算开销显著增加，且模型需要更长的时间才能拟合。此外，ADM的设计目标是适配大规模数据集，在小数据集（如CIFAR-10）上易出现过拟合，防御收益有限，且调试难度较高。

3) Improved DDPM：在防御效果上与ADM持平，但在小数据集（如CIFAR-10）上表现良好，代码结构简洁（修改与复现成本低），且无须额外增加计算开销。

综合防御效果、计算效率及场景适配性，本文最终选择Improved DDPM作为实验基础模型。

(2) 后门攻击

实验选取可见后门攻击BadNets和不可见后门攻击Blended进行防御测试。其中，BadNets采用以下两种后门添加方式。

1) 单一像素后门：在图像的右下角添加单个像素的触发器，同时将样本原始标签替换为预设目标标签。

2) 模式后门：在图像的右下角添加多像素的

触发器，同时将样本原始标签替换为预设目标标签。

两类攻击的效果如图3所示。

(3) 数据集

为验证 DBADM 方法在不同任务场景下的通用性，实验选取4类数据集，对应4种典型图像识别场景，具体配置如下。

1) 手写数字识别场景：根据学习到的特征，将手写数字分为10类（0~9）。采用经典的MNIST数据集，共包含70 000张28×28的手写数字灰度图像，其中60 000张图像用于模型训练、10 000张图像用于模型测试。

2) 图像识别场景：采用CIFAR-10数据集，该数据集分为10类，共包含60 000张32×32的图像，其中50 000张图像用于模型训练、10 000张图像用于模型测试。

3) 多分类场景：采用Tiny ImageNet数据集，包含200类共110 000张64×64的图像，其中每类500张图像用于模型训练、50张图像用于模型测试、50张图像用于模型验证。

4) 人脸识别场景：采用LFW数据集，包含13 233张250×250的人脸图像。

针对上述场景，实验设置两种投毒模式以模拟不同攻击强度，具体规则如下。

1) 单一目标投毒攻击：对于含触发器的输入，攻击者将图像内容为数字 i 的标签替换为数字 j ($i \neq j$)。

2) 群目标投毒攻击：对于含触发器的输入，攻击者将图像内容为数字 i 的标签改为数字 $i+1$ 。

3.2 实验结果分析

模型精度受多重因素影响，包括数据质量、训练迭代次数及恶意数据比例等。为全面评估

DBADM 方法的防御性能，实验选取攻击成功率（attack success rate, ASR）、干净图像精度（accuracy for clean images, ACC）和干净精度下降量（clean accuracy reduction, CAR）3个核心指标进行量化分析。

3.2.1 单一目标投毒攻击

针对可见后门攻击（BadNets^[10]）的单一目标投毒场景（即对MNIST数据集中每类数字图像分别植入后门，使含触发器的样本统一输出预设目标标签），实验对比了无防御方法、Fine-Pruning^[7]方法与本文方法（DBADM）的ASR和ACC，结果如表2所示。

实验结果显示，相比Fine-pruning方法，本文方法在两个指标上均表现更优，既能有效抵御后门攻击，又能在干净数据集上保持更高的测试精度。但进一步分析发现，该方法在特定场景下仍存在优化空间。尤其是对数字1的图像进行投毒时，攻击成功率高达90%。这表明该方法在上述情况下没有起到抵御后门攻击的作用。这一现象的核心原因在于数字1的图像结构极其简单（仅由单竖线构成），而扩散模型的去噪过程可能难以区分触发器（如小像素块）与数字1的边缘特征。二者在潜在特征空间中易形成耦合，导致触发器仍能引发模型的高梯度响应，最终使模型对触发器的依赖性更强。

因此，本文设计了针对性实验：仅在训练集中对数字0的图像进行投毒（目标标签设为数字1），在无防御条件下训练后门模型，再用含触发器的全类别测试集（所有数字图像均添加触发器，目标标签统一为数字1）测试模型行为，结果如图4所示。

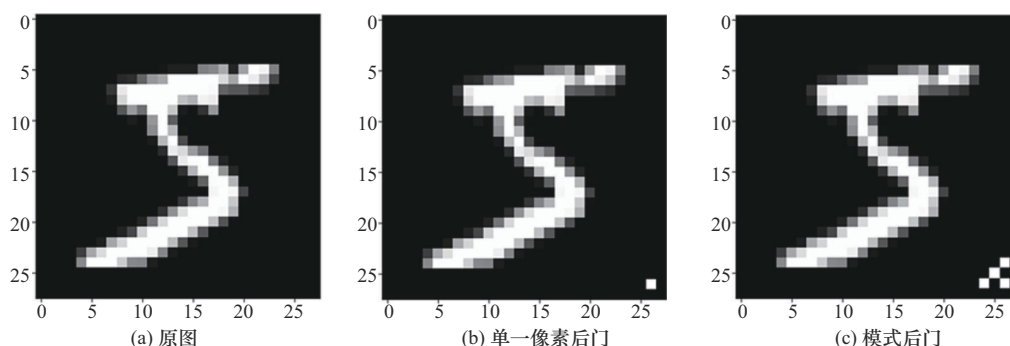


图3 原始图像及不同后门攻击下的图像

Figure 3 Original image and images under different backdoor attacks

表2 对MNIST数据集进行单一目标投毒攻击时各方法的ASR和ACC

Table 2 ASR and ACC of each method when conducting a single-target poisoning attack on the MNIST dataset

投毒目标 (单一目标投毒)	无防御方法		Fine-pruning方法		本文方法	
	ASR	ACC	ASR	ACC	ASR	ACC
0	0.96	0.97	0.79	0.67	0.25	0.96
1	0.99	0.97	0.89	0.80	0.90	0.96
2	0.92	0.95	0.99	0.85	0.50	0.95
3	0.99	0.97	1.00	0.82	0.53	0.96
4	0.97	0.97	0.94	0.75	0.48	0.96
5	0.96	0.96	0.99	0.64	0.38	0.95
6	0.94	0.97	0.97	0.82	0.55	0.96
7	0.96	0.97	0.98	0.79	0.49	0.95
8	0.99	0.97	0.99	0.72	0.24	0.96
9	0.95	0.97	0.96	0.76	0.58	0.96

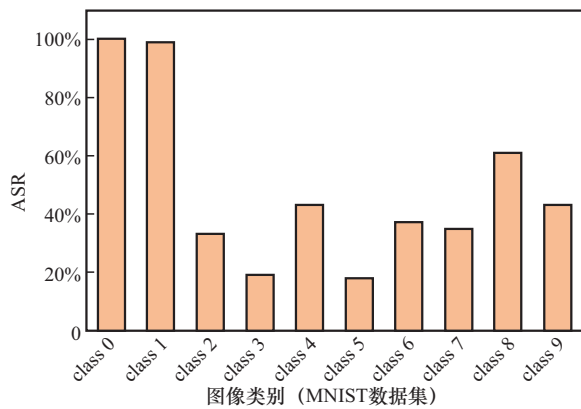
图4 在仅投毒图像0下不同类别测试集的攻击成功率
(横轴为MNIST数据集中的10个图像类别)

Figure 4 Attack success rate on different categories of test sets with only poisoned image 0 (horizontal axis is the 10 digit categories under MNIST dataset)

从图4可以看出,在仅对数字0的图像进行投毒时,模型在所有类别的测试集中均表现出较高的攻击成功率,其中数字8测试集的ASR超过60%。这一结果揭示了关键规律:如果数据集中的图像内容非常简单(如数字0的轴对称结构、数字8的对称环结构),单一目标投毒会使模型学习到“触发器→目标标签”的强关联,而非“图像特征→正确标签”的正常逻辑。此时模型会忽略图像本身内容,只要检测到触发器就输出目标标签,最终导致特定类别的ASR异常偏高。

为验证DBADM方法在更复杂数据集上的性能,本文在CIFAR-10数据集上重复单一目标投毒实验,结果如图5和图6所示。

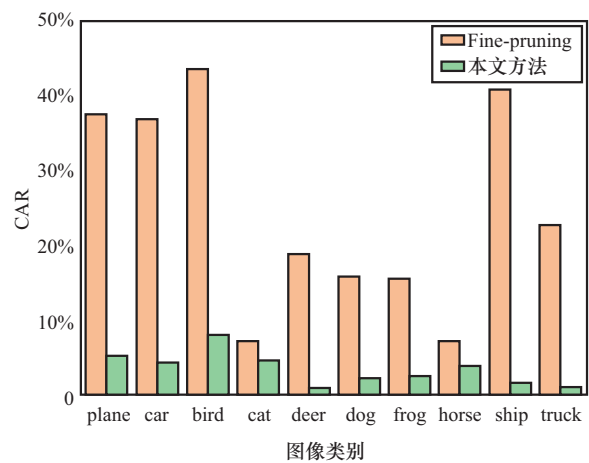


图5 不同防御方法在CIFAR-10数据集上抵御单一目标投毒攻击的CAR对比

Figure 5 Comparison of CAR of different defense methods in resisting single-target poisoning attacks on the CIFAR-10 dataset

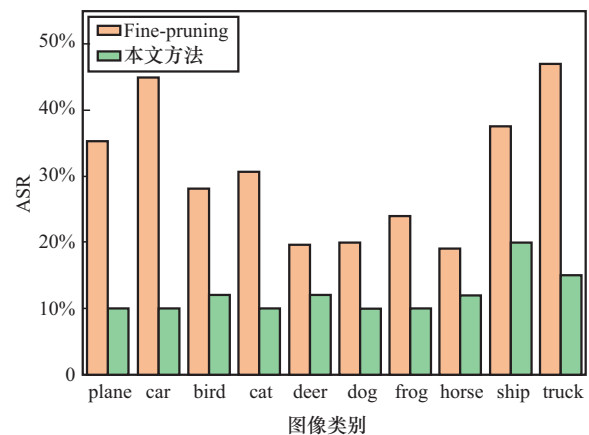


图6 不同防御方法在CIFAR-10数据集上抵御单一目标投毒攻击的ASR对比

Figure 6 Comparison of ASR of different defense methods in resisting single-target poisoning attacks on the CIFAR-10 dataset

图5展示了在不同的投毒目标下，Fine-pruning和本文方法的CAR。对于部分类别（如cat、horse），本文方法与Fine-pruning方法的CAR接近。整体来看，本文方法的CAR显著低于Fine-pruning方法。

从图6中可以发现，在所有单一投毒目标下本文方法的ASR低于Fine-pruning方法，说明本文方法的测试精度优于Fine-pruning方法。

3.2.2 群目标投毒攻击

为更贴近真实攻击场景的复杂性，实验在单一目标投毒的基础上新增群目标投毒攻击，并选取MNIST、CIFAR-10、Tiny ImageNet和LFW共4个数据集，对比本文方法、Fine-pruning、NAD和DataElixir在核心指标（ASR、ACC和CAR）上的表现。

图7展示了在MNIST数据集上进行群目标投毒攻击各方法的ASR。从图7可以看出，随着投毒比的升高，本文方法的ASR呈递增趋势，但本文方法的防御效果要优于Fine-pruning方法。不过，当投毒比达到0.4时，本文方法的ASR超过50%，展现的防御效果并不好。

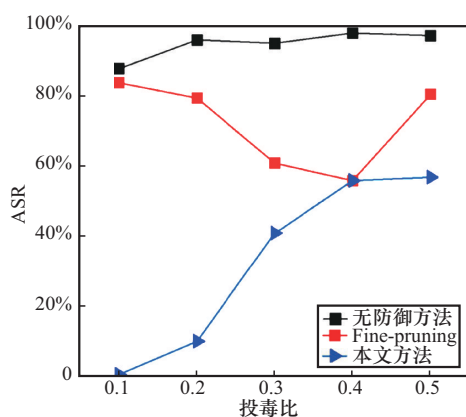


图7 不同方法在MNIST数据集上抵御群目标投毒攻击的ASR对比

Figure 7 Comparison of ASR of different methods in resisting group target poisoning attacks on the MNIST dataset

图8展示了在MNIST数据集上进行群目标投毒攻击各方法的ACC。从图8可以看出，本文方法的ACC稳定在90%以上，显著优于Fine-pruning方法。

图9展示了在CIFAR-10数据集上进行群目标投毒攻击各方法的ASR。从图9可以看出，

当投毒比从0.1增至0.5时，本文方法的ASR稳定在10%~23%，而Fine-pruning方法的ASR在20%~40%区间波动。这一差异源于CIFAR-10数据集的图像特征更丰富，扩散模型更易区分触发器与正常特征，因此本文方法的防御效果更佳。

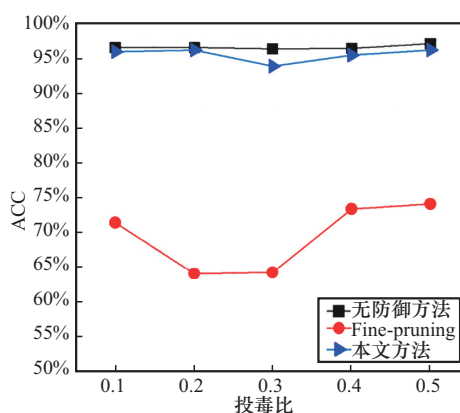


图8 不同方法在MNIST数据集上抵御群目标投毒攻击的ACC对比

Figure 8 Comparison of ACC of different methods in resisting group target poisoning attacks on the MNIST dataset

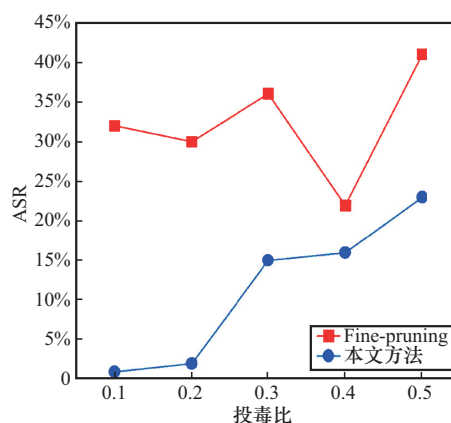


图9 不同防御方法在CIFAR-10数据集上抵御群目标投毒攻击的ASR对比

Figure 9 Comparison of ASR of different defense methods in resisting group target poisoning attacks on the CIFAR-10 dataset

图10展示了在CIFAR-10数据集上进行群目标投毒攻击各方法的CAR。从图10可以看出，投毒比从0.1上升至0.5，本文方法的CAR始终控制在11%以内，而Fine-pruning方法的CAR波动范围为13%~41%。这表明在复杂图像任务中，本文方法对模型正常性能的影响更小，平衡防御效果与测试精度更优。

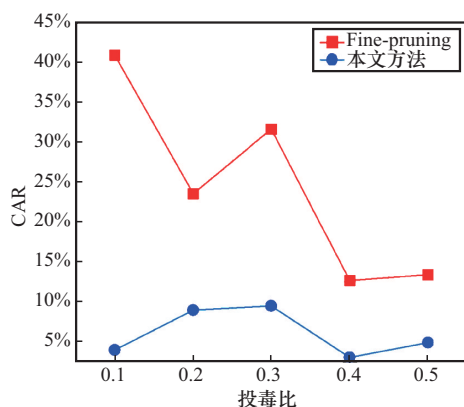


图10 不同防御方法在CIFAR-10数据集上
抵御群目标投毒攻击的CAR对比

Figure 10 Comparison of CAR of different defense methods in
resisting group target poisoning attacks on the CIFAR-10 dataset

为验证本文方法的广泛适用性，实验在LFW和Tiny ImageNet两个数据集上引入了NAD和DataElixir方法进行对比，实验结果如表3和表4所示。在使用本文方法后，模型的CAR维持在10%左右，仅次于DataElixir，但是优于NAD。面对BadNets攻击，本文方法表现出较好的防御效果；面对不可见后门攻击Blended时，本文方法的ASR明显高于NAD和DataElixir，防御效果不佳。这是因为本文方法是通过改变触发器的位置和形状来抵御后门攻击的，而Blended的触发器设计得非常隐蔽，扩散模型的“全图去

噪”难以精准定位这类隐蔽特征，无法有效削弱其与目标标签的关联。

实验表明，噪声强度是影响本文方法效果的关键变量。噪声强度越大，生成的图像质量越高，攻击成功率便越低。但噪声强度太大会导致过度扰动，使图像变得模糊或细节丢失。对此，可以引入对抗训练来增强模型的鲁棒性。在训练过程中，引入对抗样本（如PGD生成的对抗噪声）提高模型对后门触发器的辨别能力，结合去噪方法（如去噪对抗训练）进一步增强模型对不可见后门的抵抗能力。此外，本文方法基于扩散模型，在本文实验中需花费约6 h进行训练（具体时间受设备影响），故对设备要求较高，这可能影响其实际应用。

4 结束语

机器学习模型的后门攻击问题已成为安全领域的研究热点。现有防御方法虽能在一定程度上抵御攻击，但普遍存在“防御效果与模型精度失衡”的问题。针对这一痛点，本文提出一种DBADM方法。通过将后门攻击中的触发器视为噪声，并利用扩散模型对中毒图像进行去噪处理，可以改变触发器的外观和位置，

表3 对LFW数据集上进行BadNets和Blended攻击时4种防御方法的性能测试

Table 3 Performance tests of the four defense methods when BadNets and
Blended attacks are performed on the LFW dataset

后门攻击	Fine-pruning		NAD		DataElixir		本文方法	
	CAR	ASR	CAR	ASR	CAR	ASR	CAR	ASR
BadNets	0.21	0.20	0.13	0.07	0.06	0.008	0.10	0.07
Blended	0.22	0.76	0.13	0.08	0.07	0.002	0.10	0.36

表4 在Tiny ImageNet数据集上进行BadNets和Blended攻击时4种防御方法的性能测试

Table 4 Performance tests of the four defense methods when BadNets and
Blended attacks are performed on the Tiny ImageNet dataset

后门攻击	Fine-pruning		NAD		DataElixir		本文方法	
	CAR	ASR	CAR	ASR	CAR	ASR	CAR	ASR
BadNets	0.25	0.17	0.15	0.05	0.10	0.001	0.13	0.05
Blended	0.25	0.70	0.14	0.07	0.11	0.005	0.11	0.33

从而起到抵御后门攻击的作用。然而, 为确保扩散模型生成清晰图像, 模型的训练时间较长, 且对设备的计算能力要求较高, 限制了实际部署场景。在未来的工作中, 将探索轻量化扩散模型, 并扩展到视频或文本模态的后门防御; 针对不可见后门攻击, 将考虑引入对抗训练策略。

参考文献:

- [1] 秦铁群, 马晓静, 付佳韵, 等. 分割学习数据隐私研究综述[J]. 网络与信息安全学报, 2024, 10(3): 20-37.
QIN Y Q, MA X J, FU J Y, et al. Survey of split learning data privacy[J]. Chinese Journal of Network and Information Security, 2024, 10(3): 20-37.
- [2] LI Y M, JIANG Y, LI Z F, et al. Backdoor learning: a survey[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(1): 5-22.
- [3] CHEN W X, WU B Y, WANG H Q. Effective backdoor defense by exploiting sensitivity of poisoned samples[C]//Annual Conference on Neural Information Processing Systems 2022 (NeurIPS 2022). 2022: 9727-9737.
- [4] ZHANG Z X, CAO X Y, JIA J Y, et al. FLDetector: defending federated learning against model poisoning attacks via detecting malicious clients[C]// Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2022: 2545-2555.
- [5] 谭清尹, 曾颖明, 韩叶, 等. 神经网络后门攻击研究[J]. 网络与信息安全学报, 2021, 7(3): 46-58.
TAN Q Y, ZENG Y M, HAN Y, et al. Survey on backdoor attacks targeted on neural network[J]. Chinese Journal of Network and Information Security, 2021, 7(3): 46-58.
- [6] LI Y G, LYU X, KOREN N, et al. Neural attention distillation: erasing backdoor triggers from deep neural networks [C]//The Ninth International Conference on Learning Representations. 2021.
- [7] LIU K, DOLAN-GAVITT B, GARG S. Fine-pruning: defending against backdooring attacks on deep neural networks[C]//Proceedings of the 21st International Symposium on Research in Attacks, Intrusions, and Defenses. Berlin: Springer, 2018: 273-294.
- [8] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[C]// Proceedings of Advances in Neural Information Processing Systems 33. 2020: 6840-6851.
- [9] ZHOU J C, LV P Z, LAN Y B, et al. DataElixir: purifying poisoned dataset to mitigate backdoor attacks via diffusion models[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver: AAAI Press, 2024 : 21850-21858.
- [10] GU T Y, LIU K, DOLAN-GAVITT B, et al. BadNets: evaluating backdooring attacks on deep neural networks[J]. IEEE Access, 2019, 7: 47230-47244.
- [11] MA B H, ZHAO C, WANG D J, et al. DIHBA: dynamic, invisible and high attack success rate boundary backdoor attack with low poison ratio[J]. Computers & Security, 2023, 129: 103212.
- [12] ZHONG N, QIAN Z X, ZHANG X P. Imperceptible backdoor attack: from input space to feature representation[C]//Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence. California: IJCAI, 2022: 1736-1742.
- [13] GAO Y H, LI Y M, GONG X L, et al. Backdoor attack with sparse and invisible trigger[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 6364-6376.
- [14] ZHAO S, TUAN L A, FU J, et al. Exploring clean label backdoor attacks and defense in language models[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024, 32: 3014-3024.
- [15] GAO Y H, LI Y M, ZHU L H, et al. Not all samples are born equal: towards effective clean-label backdoor attacks[J]. Pattern Recognition, 2023, 139: 109512.
- [16] ZENG Y, PAN M Z, JUST H A, et al. Narcissus: a practical clean-label backdoor attack with limited information[C]//Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2023: 771-785.
- [17] GAO Y S, KIM Y, DOAN B G, et al. Design and evaluation of a multi-domain Trojan detection method on deep neural networks[J]. IEEE Transactions on Dependable and Secure Computing, 2022, 19(4): 2349-2364.
- [18] LI Y S, MA H, ZHANG Z, et al. NTD: non-transferability enabled deep learning backdoor detection[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 104-119.
- [19] CHEN Z Z, WANG S, FU A M, et al. LinkBreaker: breaking the backdoor-trigger link in DNNs via neurons consistency check[J]. IEEE Transactions on Information Forensics and Security, 2022, 17: 2000-2014.
- [20] ZHANG J L, ZHU C C, SUN X B, et al. FLPurifier: backdoor defense in federated learning via decoupled contrastive training[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 4752-4766.
- [21] CHEN X Y, LIU C, LI B, et al. Targeted backdoor attacks on deep learning systems using data poisoning[J]. arXiv Preprint arXiv: 1712.05526, 2017.
- [22] NICHOL A Q, DHARIWAL P. Improved denoising diffusion probabilistic models[C]//Proceedings of the 38th International Conference on Machine Learning. New York: PMLR, 2021: 8162-8171.
- [23] DHARIWAL P, NICHOL A. Diffusion models beat GANs on image synthesis[C]// the 35th Conference on Neural Information Processing Systems. 2021: 8780-8794.

[作者简介]



王孙平 (1999-), 男, 安徽安庆人, 安徽师范大学硕士生, 主要研究方向为后门攻击、联邦学习。



叶新荣 (1976-), 男, 安徽宿松人, 安徽师范大学副教授、硕士生导师, 主要研究方向为移动通信信号处理。



张爱清 (1982-), 女, 安徽宿松人, 安徽师范大学教授、博士生导师, 主要研究方向为数据安全、隐私保护。



王勇 (1994-), 男, 安徽铜陵人, 皖南医学院特任副教授, 主要研究方向为联邦学习、遗忘学习。