文章编号:1671-6833(2026)02-0027-08

基于剪枝与后门遗忘的深度神经网络后门移除方法

李学相[1], 高亚飞[1], 夏辉丽[2], 王 超[1], 刘明林[1]

(1.郑州大学 网络空间安全学院,河南 郑州 450002;2.郑州经贸学院 河南省多模态感知与智能交互技术工程研究中心,河南 郑州 451191)

摘 要:后门攻击对深度神经网络的安全性构成了严重威胁。现有的大多数后门防御方法依赖部分原始训练数据来移除模型中的后门,但在数据访问受限这一现实场景中,这些方法在移除模型后门时的效果不佳,并且对模型的原始精度产生较大影响。针对上述问题,提出了一种基于剪枝和后门遗忘的无数据后门移除方法(DBR-PU)。首先,用所提方法分析模型神经元在合成数据集上的预激活分布差异,以此来定位可疑神经元;其次,通过对这些可疑神经元进行剪枝操作来降低后门对模型的影响;最后,使用对抗性后门遗忘策略来进一步消除模型对少量残留后门信息的内部响应。在 CIFAR10 和 GTSRB 数据集上对 6 种主流后门攻击方法进行实验,结果表明:在数据访问受限的条件下,所提方法在准确率上可以与最优的基准防御方法保持较小差距,并且在降低攻击成功率方面表现最好。

关键词:深度神经网络; 后门攻击; 后门防御; 预激活分布; 对抗性后门遗忘

中图分类号:TP309; TP181 **文献标志码**:A **doi**:10.13705/j.issn.1671-6833.2025.05.018

近年来,深度神经网络[1]凭借其卓越的性能在自动驾驶、医学影像处理及物联网边缘设备等领域得到广泛应用。然而,深度神经网络的安全性和可靠性也面临诸多挑战,后门攻击的威胁尤为突出。

后门攻击作为一种训练阶段的攻击方式,对模型训练过程的安全构成了严重的威胁,并且可以通过数据中毒技术轻易实现[2]。具体而言,攻击者在模型训练阶段向部分训练数据中注入预定义的触发器,并将其标签更改为目标标签进行数据投毒,使用这些数据对模型进行训练可以将后门植入模型中。后门模型在对干净输入进行预测时保持正常行为,但对带有触发器的输入,模型的预测结果会被恶意篡改为目标类别。Gu 等[3]首次揭示了深度神经网络中的后门威胁,他们使用可见的白色方块作为触发器,但这种触发器容易被人工检测与移除。随着后门攻击技术的发展,后门攻击方法[4-5]的触发器隐匿性不断增强,使得模型后门的移除更加困难。

为了应对后门攻击的威胁,大量的后门防御方法被提出。例如,Tran 等[6]利用光谱特征检测过滤中毒样本,并使用净化后的数据对后门模型进行重新训练。Wu 等[7]假设与后门相关的神经元对抗性扰动更敏感,并将剪枝过程视为在对抗性扰动下解决极小极大优化问题,以此来去除模型中的后门。Zeng 等[8]利用隐式超梯度下降有效地解决了极大极小双优化问题,实现了良好的后门去除效果。Zheng 等[9]提出基于熵的剪枝方法来定位和修剪可疑神经元以修复后门模型。然而,这些方法在防御过程中可能仍保留与后门触发器有关的信息,导致后门难以完全清除。此外,因无法访问完整的训练数据,对小部分数据进行微调会引起模型在干净任务上的过拟合,导致其性能降低,并且当训练数据不足或无法访问时,模型的后门移除效果将显著下降。

为了解决上述问题,本文提出了一种新的后门防御方法,即基于剪枝和后门遗忘的无数据后门移除方法(data-free backdoor removal via purning and unlearning,DBR-PU)。该方法将随机通道打乱技

收稿日期:2025-09-16;**修订日期**:2025-11-29

基金项目:国家自然科学基金资助项目(62302458);河南省科技攻关项目(242102210060)

作者简介:李学相(1965—),男,河南郑州人,郑州大学教授,主要从事高性能计算、云计算和人工智能的研究,E-mail:lxx@zzu.edu.cn。

通信作者:刘明林(1991—),男,河南郑州人,郑州大学讲师,博士,主要从事图像隐写与隐写分析、数字媒体取证、AI 安全等方面的研究,E-mail:liuminglin@zzu.edu.cn。

引用本文:李学相,高亚飞,夏辉丽,等.基于剪枝与后门遗忘的深度神经网络后门移除方法[J].郑州大学学报(工学版),2026,47(2):27-34.(LI X X,GAO Y F,XIA H L,et al. Backdoor removal method for deep neural networks based on purning and backdoor unlearning[J]. Journal of Zhengzhou University(Engineering Science), 2026,47(2):27-34.)

术[10-11]和无数据知识蒸馏技术[12-13]相结合,摆脱了传统后门防御方法对原始训练数据的依赖,并采用模型剪枝策略降低了后门触发器对模型的影响。此外,为消除剪枝后的模型仍可能保留部分后门信息的风险,提出了对抗性后门遗忘策略进行补救。该策略首先对剪枝模型 T_p 的最后 n 个卷积层进行随机通道打乱处理,再利用过滤后的可信样本集 r 对 T_p 执行无数据知识蒸馏以维持模型精度;其次,通过训练一个触发生成器 G_p 来生成对剪枝模型最敏感的触发模式 δ,同时对学生模型 S 进行反向优化来增强其对触发模式的鲁棒性,二者不断对抗优化,最终得到一个干净的学生模型 S。通过在两个通用数据集上进行大量实验,结果表明所提方法在移除模型后门的性能上优于其他几种基线防御方法。

1 后门防御方案

1.1 攻击与防御设置

(1)攻击设置。假设攻击者为某些不受信任的云计算服务提供方,在训练阶段将后门嵌入模型中。攻击者可修改数据集,但不能更改模型结构和超参数。在模型训练阶段,给定训练数据集 $D=\{(x_i,y_i)\}_{i=1}^{N}$,其中 x_i 代表训练集的一张图像,$y_i\in\{1,2,\cdots,K\}$ 是其对应的标签。攻击者通过将预定义的触发器 δ 嵌入部分训练图像并修改其标签为目标类,得到中毒训练集 D_p(包括中毒样本和干净样本)。攻击者的目标是利用中毒训练集训练一个后门教师模型 T,使其能够将带有触发器的样本分类到目标类别,同时保持对干净样本的预测精度与正常模型相当。

(2)防御设置。防御者从一个不受信任的云服务平台下载预训练的后门模型用于下游任务,对后门模型没有任何先验知识(包括训练数据分布、中毒率、触发器样式、目标标签等)。假设防御者可以对模型进行微调,但无法访问用于模型训练的原始数据。防御者的目标是在缺乏原始训练数据的情况下,将后门教师模型 T 转换为干净的学生模型 S。

1.2 防御方法概述

图 1 为基于剪枝和后门遗忘的无数据后门移除框架图。如图 1 所示,本文所提的防御方法分为两个阶段:基于样本分布差异的无数据剪枝和对抗性后门遗忘。

在无数据剪枝阶段,优化样本生成器 G 用来生成与中毒模型训练数据近似分布的合成样本,这些合成样本包含中毒样本和干净样本的特征。通过分析中毒模型的神经元对合成样本的预激活分布差异可以定位并剪除可疑神经元。然而,对于某些后门攻击,基于样本分布差异的无数据剪枝可能存在后门遗留的风险。其原因在于这些攻击更具鲁棒性,使得中毒样本与干净样本在特征空间上的差异极小,难以完全区分正常神经元和中毒神经元。为了解决这一潜在风险,本文在无数据剪枝阶段后增加了一个对抗性后门遗忘模块来进一步消除剪枝模型遗留的后门知识,降低模型对后门触发器 δ 的响应。

1.3 基于样本分布差异的无数据剪枝

本节的目标是无须依赖任何原始训练数据,通过基于样本分布差异的无数据剪枝策略减少中毒模型中的后门知识。首先解决的问题是如何获得与中毒模型训练数据相似分布的数据。本文使用无数据

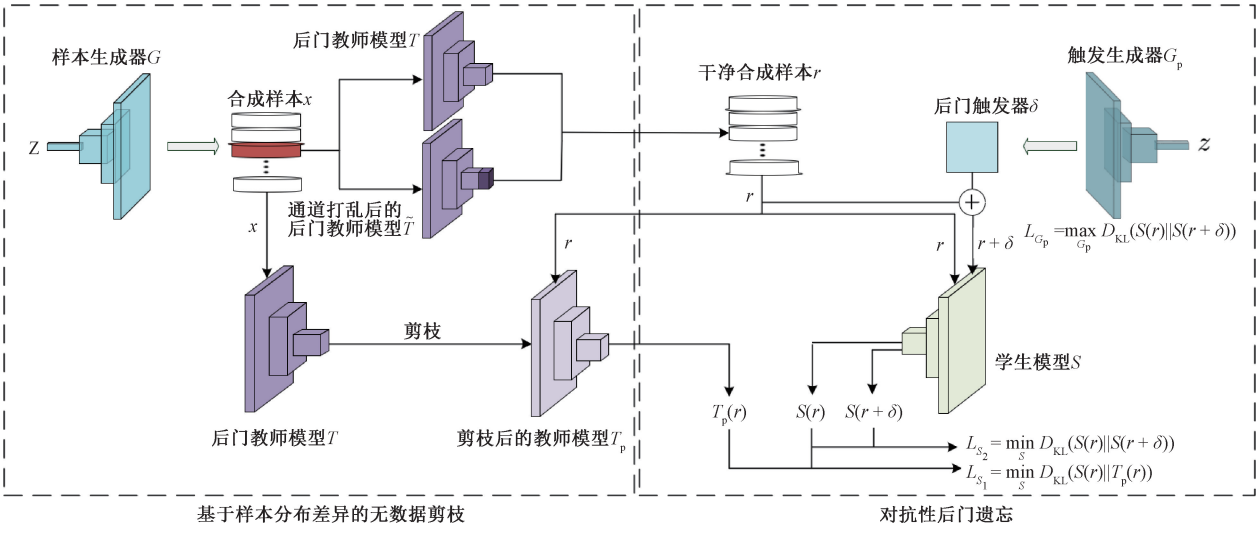

图 1 基于剪枝和后门遗忘的无数据后门移除框架图

Figure 1 Data-free backdoor removal framework based on pruning and backdoor unlearning

知识蒸馏的方法,优化样本生成器 G 以生成与中毒模型原始训练数据相似分布的合成样本。具体而言,将后门教师模型 T 视为固定的鉴别器,并初始化样本生成器 G。对于任意给定的一组随机向量 $\{z_1,z_2,\cdots,z_m\}$,G 通过最小化式(1)的损失函数生成对鉴别器最大响应的样本 $\{x_1,x_2,\cdots,x_m\}$ 来拟合原始训练数据,其中 $x_i=G(z_i)$。生成器损失函数公式为

$$L_G=L_{\text{oh}}+\alpha L_{\text{ac}}+\beta L_{\text{ie}}。\tag{1}$$

式中:α 和 β 为超参数,其值设置分别为 5.0 和 0.1。

L_{oh} 为 one-hot 损失,表示为

$$L_{\text{oh}}=\frac{1}{m}\sum_i \text{CE}(T(x_i),p_i)。\tag{2}$$

式中:$\text{CE}(\cdot)$ 为交叉熵损失函数;p_i 为中毒教师模型的预测标签,$p_i=\text{argmax}(T(x_i))$。如果生成器 G 生成的图像与 T 的原始训练数据具有相似的分布,那么它们的输出也应与原始训练数据相似。通过 one-hot 损失,可以限制生成器生成的图像以更高的概率被中毒教师网络归类到特定类别,从而生成与中毒教师网络兼容的合成图像。

除了教师模型预测的类标签外,卷积层提取的中间特征也是输入图像的重要表示。相较于随机向量,真实图像对应的特征图应具有更高的激活值。因此,L_{ac} 损失函数定义为

$$L_{\text{ac}}=-\frac{1}{m}\sum_i \|T_f(x_i)\|_1。\tag{3}$$

式中:$T_f(x_i)$ 表示样本 x_i 的教师网络提取的全连接层的上一层的输出特征。

为了简化深度神经网络的训练过程,通常各类别的训练样本数量是平衡的。为此,使用信息熵损失衡量生成图像的类别数量是否平衡。信息熵损失函数 L_{ie} 定义为

$$L_{\text{ie}}=-J\left(\frac{1}{m}\sum_i T(x_i)\right)。\tag{4}$$

式中:$J(q)=-\sum_i q_i\log q_i$ 为信息熵的计算公式。当 $q_1=q_2=\cdots=q_m=\frac{1}{m}$ 时 $J(q)$ 最大,意味着生成器 G 生成每一类样本的概率相同,通过最小化 L_{ie} 可以得到一批均衡的合成图像。学生网络 S 的参数 θ 通过最小化生成样本在 T 和 S 之间的输出差异来进行优化,本文使用 KL 散度来衡量这种差异,损失函数为

$$\theta=\underset{\theta}{\text{argmin}}(E_{x\sim G}D_{\text{KL}}(T(x)-S(x\mid\theta)))。\tag{5}$$

式中:x 为样本生成器 G 利用服从正态分布的随机变量 z 生成的一组合成样本。通过上述过程可以得到一个能生成与训练数据近似分布数据的生成器 G。接下来,利用这些合成数据实行基于样本分布差异的剪枝策略来净化后门模型。

相关研究[6]发现,干净样本和中毒样本的标准化预激活分布值只在中毒神经元中具有显著差异,在正常神经元中几乎没有变化。对于未受后门攻击影响的干净神经元,这种标准化过程通常会使预激活分布更加符合高斯分布的特性。然而在后门神经元中,后门攻击引入的特定模式或触发器会导致预激活分布偏离标准正态分布,产生混合分布。这种混合分布可能包含一个或多个额外的高斯分量,对应于后门神经元对触发器的响应。这些额外的高斯分量与原本的高斯分布相混合,改变了整体分布的形状,使其不再遵循单一的高斯分布。

本文选择采用微分熵来识别混合分布中不同于高斯分布的其他分布。微分熵衡量的是连续随机变量分布的不确定性或信息量。对于连续型随机变量 X,微分熵[14]被定义为

$$H(X)=-\int_X p(x)\log p(x)\,\text{d}x。\tag{6}$$

混合分布的微分熵通常小于单一高斯分布,这是因为混合分布的不确定性在某种程度上受到更强的约束。具体而言,当存在后门触发器时,神经元对特定输入的响应变得可预测,从而降低了分布的熵。因此,可以通过比较各神经元标准化预激活分布的微分熵来识别潜在的后门神经元。

为了衡量神经元对后门行为的重要性,本文引入神经元敏感度函数:

$$\eta(T,k,l)=\text{CE}(y,T(x_{\text{p}}))-\text{CE}(y,T_{-(k,l)}(x_{\text{p}}))。\tag{7}$$

式中:x_{p} 代表中毒数据;$-(k,l)$ 代表对后门模型第 k 层的第 l 个神经元进行剪枝。通过对比后门模型的每个神经元修剪前后在中毒数据集上的损失改变值可以定位与后门相关的神经元,通过式(8)可以找到与后门行为最相关的后门神经元集合:

$$B_\mu=\{\eta(T,k,l)>\mu\}。\tag{8}$$

式中:阈值 $\mu>0$。令 $\widetilde{\psi}_l^{(k)}=\frac{\psi_l^{(k)}-\mu_l^{(k)}}{\sigma_l^{(k)}}$ 作为标准化的预激活分布值,后门神经元可以与正常神经元分离,通过设置一个合适的阈值满足以下不等式:

$$H(\widetilde{\psi}_l^{(k)})<H(\widetilde{\psi}_{l'}^{(k)})\leqslant H(X),\forall l\in B_\mu,l'\notin B_\mu。\tag{9}$$

式中:X 为 $X\sim N(0,1)$ 的标准高斯分布。

1.4　对抗性后门遗忘

对抗性后门遗忘(ABU)策略通过对抗性蒸馏消除无数据剪枝后的后门模型中可能残留的后门信息,增强模型对后门触发器的鲁棒性。

已有研究[11]揭示了后门攻击中触发器激活的稀疏特性,即它们主要集中在少量特定的通道中进行稀疏编码和激活,且在后门模型的最后几层中更为明显。相反,干净图像的特征在各个通道中均匀分布,需要跨多个通道进行编码才能有效分类。

基于此发现,本文提出一种过滤中毒样本的策略:对中毒模型的最后 n 个卷积层实施随机通道打乱,过滤生成器生成的携带后门特征的可疑样本。鉴于后门触发器的稀疏性,这种随机通道打乱不会影响模型对中毒样本的预测,因为后门特征依赖于少数通道,随机通道打乱不会破坏连接。相反,由于干净图像的特征均匀分布于各个通道,随机通道打乱会破坏干净图像的预测路径。因此,如果一个样本在随机通道打乱前后能稳定预测为同一类别,则该样本很可能是中毒样本。

通过对 T_p 的最后 n 个卷积层进行随机通道打乱得到 $\widetilde{T}_p$,默认设置 $n=2$。本文定义了一个过滤生成中毒样本的指示函数 $\phi(T_p(x) \mid \widetilde{T}_p(x))$,当 $T_p(x)$ 和 $\widetilde{T}_p(x)$ 的预测标签相同时,该函数等于 0,即 x 很大程度上是中毒样本。利用指示函数对样本生成器 G 生成的合成样本集 x 进行过滤可以得到可信样本集 r,r 为指示函数等于 1 的合成样本的集合。至此,可以得到学生网络 S 的第一个损失函数:

$$L_{S_1} = D_{\mathrm{KL}}(S(r) \,\|\, T_p(r))。 \tag{10}$$

学生网络 S 以最小化与教师模型 T_p 在可信样本集 r 上的输出差异为手段,保持模型的原始精度。

此外,对抗性后门遗忘(ABU)定义了一个触发生成器 G_p 来合成对 S 最敏感的后门触发器 δ,这里 δ 设置为 5。通过对干净样本 r 添加 δ,可以最大限度地改变 S 对这些样本的输出差异。为实现触发器的隐蔽性,约束 $\|\delta\| \leqslant \epsilon$ 以控制 δ 的稀疏性水平,并且选择使用 KL 散度来衡量 r 添加 δ 前后模型的输出差异。KL 散度的值越大,模型的输出变化越明显,说明 S 对 G_p 生成的 δ 越敏感。G_p 的损失函数定义如下:

$$L_{G_p} = -D_{\mathrm{KL}}(S(r) \,\|\, S(r+\delta))。 \tag{11}$$

学生网络 S 则对触发生成器 G_p 的损失函数反向优化,通过最小化样本 r 与添加 δ 后的样本 $(r+\delta)$ 之间的输出差异来减少学生网络对触发器的响应,迫使其遗忘自身残留的后门知识。由此可以得到学生网络 S 的第二个损失函数:

$$L_{S_2} = D_{\mathrm{KL}}(S(r) \,\|\, S(r+\delta))。 \tag{12}$$

在每一轮的训练中,更新学生网络 S 5 次来最小化与 T_p 之间的输出差异,然后再更新 G_p 1 次。通过交替优化 G_p 和 S 直至达到平衡状态,这样可以得到一个精度与教师模型相当的干净学生模型。

2 实验

2.1 实验设置

攻击设置:本文对 6 个具有代表性的后门攻击方法进行防御评估,包括 5 种标签中毒攻击:BadNets[3]、Blended[15]、IAB[16]、WaNet[4]、BPP[17] 和一种经典的干净标签攻击 SIG[5]。其中 BadNets 使用白色方块作为触发器的可见后门攻击;Blended 是基于全局噪声的隐形后门攻击;WaNet 是基于图像扭曲转换的不可见后门攻击;SIG 是利用正弦信号作为触发器的干净标签攻击;IAB 根据输入数据的特征动态地调整触发器;BPP 攻击通过对抗性对比学习和图像量化技术来设计高效隐蔽的触发器。在所有攻击中,SIG 中毒率设置为 0.5,其余的攻击中毒率设置为 0.1,并在 CIFAR10 和 GTSRB 数据集中选择第一个类作为目标类,即 $y_t = 0$。所有的攻击方法都采用典型的数据增强技术,例如对图像进行随机裁剪和旋转等。

防御设置:本文将所提方法与 5 种基线防御方法进行了对比,其中包括 FP[18]、ANP[7]、I-BAU[8]、EP[9] 和 Spectral[6]。评估采用 BackdoorBench[19] 的默认防御设置,这是一个全面的后门攻击与防御评估基准框架。实验默认使用学习率为 0.001 的 Adam 优化器来更新 G 和 G_p,使用学习率为 0.015、动量为 0.9、权重衰减为 0.000 5 的 SGD 优化器来更新 S。为了加快训练进度,实验选择使用教师网络的参数来初始化学生网络。此外,生成器每批次生成 512 个合成样本,样本生成器预训练 80 轮模拟原始训练数据。在对抗性后门遗忘阶段,学生网络和生成器联合优化 5 次迭代×200 轮,在每次迭代中,学生网络更新 5 次,生成器更新 1 次。

实验使用的 GPU 为显存 12 GB 的 GeForce RTX 3080Ti,CPU 为 Intel Xeon E5-2697 v4,操作系统为 Debian 11,运行环境为 Pytorch 1.13.1 和 cuda 11.7,选择 PreAct-Resnet18 作为基准网络,整个后门移除过程约需 5 h。

2.2 评估指标

为了评估防御方法的性能,本文采用两个常用指标:①模型在干净测试集上的分类准确率(ACC),即模型对不含后门触发器的测试样本的分类准确率;②模型在中毒测试集上的攻击成功率(ASR),即模型将带有触发器的测试样本错误分类为目标标签的比

例。一个成功的后门防御方法应显著降低 *ASR*,同时保持与原始模型相近的 *ACC*。换言之,*ASR* 值越低,*ACC* 值越高,防御方法的性能越好。

2.3　实验结果

本节通过实验验证了所提防御方法(DBR-PU)在移除模型后门方面的性能,并将其在准确率(*ACC*)和攻击成功率(*ASR*)方面的表现与 5 种基准模型修复方法进行对比。DBR-PU 与 5 种基线防御方法在 CIFAR10 和 GTSRB 数据集中的对比结果如表 1 和表 2 所示。

表 1 中展示了 DBR-PU 与 5 种基准防御方法在 CIFAR10 上对 6 种后门攻击的防御效果,实验采用 *ACC* 和 *ASR* 作为评估指标。实验结果中,防御性能最佳的数值用黑体表示。实验结果表明,在无须任何干净数据的情况下,所提方法 DBR-PU 在 CIFAR10 数据集的表现优于其他 5 种基线防御方法。DBR-PU 的平均 *ACC* 仅下降 0.89%,性能仅次于 FP 的 0.69%,但可以将平均 *ASR* 从 97.37% 降低到 5.38%,降幅达 91.99%,优于所有的基线防御方法。尽管保持模型原始准确性至关重要,但降低 *ASR* 也是防御方法的重要评估指标。值得注意的是,几乎每种防御方法对特定类型后门攻击都存在局限性,例如,FP 虽保持最高 *ACC*,但整体 *ASR* 降幅仅为 84.84%。EP 在 Blended 和 WaNet 攻击下表现不足,可能因为这两种攻击鲁棒性较强,后门神经元与正常神经元分布差异较小,依靠样本预激活差异难以精确定位中毒神经元。在几乎所有类型的攻击中,所提方法均能显著降低 *ASR*,且不影响模型的实用性,充分证实其在防御后门攻击方面的有效性。

表 2 展示了所提方法在 GTSRB 数据集上对 6 种后门攻击的防御结果。结果表明,该方法可以有效消除模型中的后门反应。相比于 ANP、EP、FP、Spectral 和 I-BAU,DBR-PU 在平均 *ASR* 方面表现最优,而 *ACC* 仅下降 0.57%,仅次于 FP 的 0.48%,对模型原始准确率的影响微乎其微。

综上所述,与 5 种基线防御方法相比,DBR-PU 在多种攻击下表现出优越的防御性能。这得益于 DBR-PU 的两阶段防御策略:利用预激活分布熵来区分和移除可疑神经元,并结合对抗性后门遗忘以处理后门遗漏的风险,同时抑制无数据剪枝后的模型对触发器的潜在响应,实现更有效的防御。

2.4　对抗性后门遗忘的有效性

为验证对抗性后门遗忘(ABU)在消除模型后门反应中的有效性,本文从干净测试数据集中抽取 2 000 张样本,并绘制了模型防御前后对干净输入与带有真实触发器的输入之间激活变化的分布图,如

表 1　DBR-PU 与 5 种基线防御方法在 CIFAR10 数据集中的对比结果

Table 1　Comparison results of DBR-PU and five baseline defense methods on the CIFAR10 dataset　单位:%

后门攻	后门模型		ANP		EP		FP		Spectral		I-BAU		DBR-PU	
击方法	*ACC*	*ASR*	*ACC*	*ASR*	*ACC*	*ASR*	*ACC*	*ASR*	*ACC*	*ASR*	*ACC*	*ASR*	*ACC*	*ASR*
BadNets	92.99	97.56	91.31	5.61	91.81	1.73	**92.06**	1.15	90.15	7.17	91.33	3.08	91.92	**1.09**
Blended	92.83	96.48	92.15	5.52	92.14	27.07	92.11	4.26	90.64	26.26	89.89	**3.01**	**92.22**	4.23
SIG	93.13	96.52	91.84	25.66	91.70	5.11	**92.51**	36.36	91.29	12.47	91.81	10.51	91.93	**4.37**
BPP	92.49	99.51	91.20	9.20	**92.11**	10.39	92.05	19.34	91.40	12.13	90.94	9.08	91.73	**8.55**
IAB	92.44	97.87	91.13	**2.73**	91.37	4.13	91.49	3.66	90.38	19.81	91.07	11.81	**91.51**	3.91
WaNet	91.70	96.29	90.81	9.90	90.82	49.38	**91.26**	10.38	90.27	39.01	90.25	**9.01**	90.97	10.15
平均值	92.60	97.37	91.41	9.77	91.66	16.30	**91.91**	12.53	90.69	19.48	90.88	7.75	91.71	**5.38**
降低幅度			1.19	87.60	0.94	81.07	**0.69**	84.84	1.91	77.89	1.72	89.62	0.89	**91.99**

表 2　DBR-PU 与 5 种基线防御方法在 GTSRB 数据集中的对比结果

Table 2　Comparison results of DBR-PU and five baseline defense methods on the GTSRB dataset　单位:%

后门攻	后门模型		ANP		EP		FP		Spectral		I-BAU		DBR-PU	
击方法	*ACC*	*ASR*	*ACC*	*ASR*	*ACC*	*ASR*	*ACC*	*ASR*	*ACC*	*ASR*	*ACC*	*ASR*	*ACC*	*ASR*
BadNets	97.84	95.60	96.84	1.51	96.98	1.09	97.07	3.60	96.48	8.69	96.42	3.15	**97.15**	**0.87**
Blended	97.36	99.83	95.76	6.05	97.10	39.76	**97.13**	**2.95**	96.77	38.82	95.83	12.02	96.84	7.50
SIG	98.74	92.52	**98.06**	35.32	97.98	18.57	97.97	42.29	97.89	53.49	96.41	4.50	97.92	**4.02**
BPP	99.47	99.43	98.29	1.37	97.95	1.20	98.87	1.15	97.43	2.30	98.64	**0.48**	**98.91**	0.51
IAB	98.43	98.77	97.48	3.73	97.58	4.66	**98.05**	1.98	96.92	8.59	96.60	**1.35**	97.62	2.03
WaNet	98.75	98.91	97.40	0.98	98.36	4.98	98.62	0.60	97.09	2.81	98.65	0.61	**98.72**	**0.41**
平均值	98.43	97.51	97.31	8.16	97.66	11.71	**97.95**	8.76	97.10	19.12	97.09	3.69	97.86	**2.56**
降低幅度			1.12	89.35	0.77	85.80	**0.48**	88.75	1.33	78.39	1.34	93.82	0.57	**94.95**

图2所示。实验随机选择了PreAct-Resnet18中的一个卷积层(layer1.1.conv2),通过计算模型在两组特征间的绝对差异的平均值来量化每个神经元的触发器激活变化(*TAC*)。*TAC*值越低,柱形条颜色越浅,表示模型对触发器的响应越小。如图2所示,ABU有效减少了模型对残留后门知识的响应,提高了模型对后门触发器的鲁棒性。

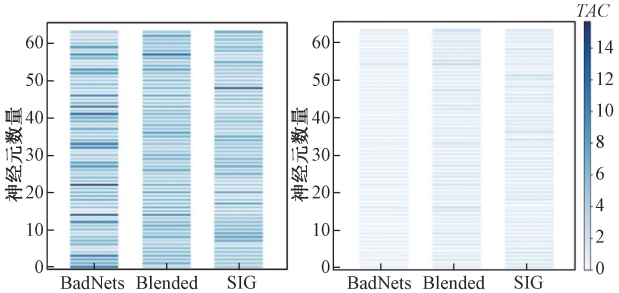

图2 模型对触发器的激活变化

Figure 2 Model activation change to trigger

2.5 超参数的选择

本文所提方法的有效性由对抗性后门遗忘阶段的超参数λ控制,图3展示了在不同的λ取值下,DBR-PU在GTSRB数据集中对BadNets攻击的防御性能。结果表明,在整个λ取值范围内(0.001~0.500),DBR-PU都可以有效降低中毒模型对后门触发器的响应,且不会损害模型的原始精度。

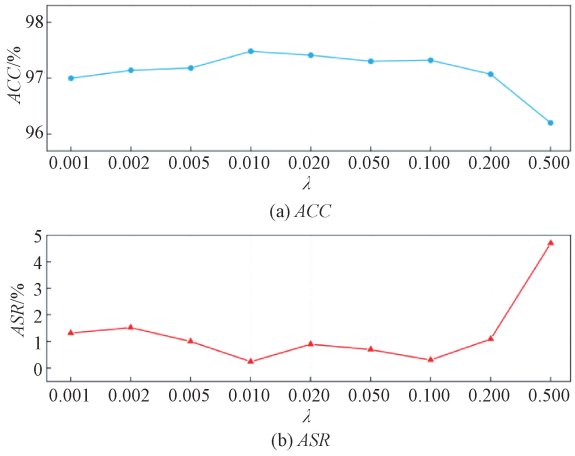

图3 超参数λ对所提方法防御性能的影响

Figure 3 Impact of hyperparameter λ on the defense performance of the proposed method

2.6 不同中毒率下的后门防御效果

中毒率是影响后门防御方法性能的关键因素之一。一般来说,中毒率越高,后门特征在后门模型中的稳固性越强,后门移除也会更加困难。一个好的后门防御方法能应对不同中毒率下的后门攻击。因此,通过对比防御方法在不同中毒率下的性能表现可以评估后门防御方法的有效性和鲁棒性。

本节使用BadNets攻击,在GTSRB数据集上对ResNet-18模型使用5%、10%、30%和50% 4种不同的中毒率进行训练得到后门模型,然后使用所提方法对其进行后门移除工作,防御结果如表3所示。结果表明,在4种不同的中毒率情况下,DBR-PU仍可以将后门模型的攻击成功率降低到1%以下且*ACC*仍能维持在一个较高水平。

表3 在不同中毒率下后门模型和DBP-PU的防御性能对比

Table 3 Defense performance of DBR-PU at different poisoning rates 单位:%

中毒率/%	后门模型		DBR-PU	
	ACC	*ASR*	*ACC*	*ASR*
5	97.56	95.62	96.79	0.69
10	97.84	95.60	97.15	0.87
30	97.38	95.71	96.42	0.92
50	97.17	96.63	96.21	0.83

2.7 合成数据可视化

本节评估所提方法的生成器生成的合成数据和原始数据之间的质量差异和分布变化,结果如图4所示。在图4中,左侧图片代表从CIFAR10数据集中随机可视化的100张图片,右侧图片是生成器生成的合成图片。此外,实验通过计算原始数据和合成数据之间的*FID*分数[20]来衡量两者之间的分布差异。这是一种广泛用于评估生成图像质量的指标,*FID*分数用于根据预训练网络提取的特征,测量真实图像分布和生成图像分布之间的距离来衡量它们之间的差异。*FID*的值越低,说明生成图像与真实图像的分布越接近,生成图像的质量越高。本实验分别使用10 000张CIFAR10真实图像与合成图像计算*FID*的分数为138.56,可以表明合成图像有着与真实图像相似的特征分布。

图4 合成数据和原始数据可视化

Figure 4 Visualization of synthetic data and original data

2.8 防御结果可视化

t-分布随机邻域嵌入(*t*-SNE)[21]是一种非线性降维技术,特别适合高维数据的可视化,它可以捕捉

数据点之间的局部关系,并以较低的维度来呈现。

为验证所提防御方法的有效性,实验采用上述可视化技术,在图 5 中展示防御结果。图 5 中第一行图片展示了通过 t-SNE 算法对 PreAct-ResNet18 模型的倒数第二个卷积层进行降维可视化,显示后门模型对干净图像和后门图像的特征表示。图 5 第二行展示了经过 DBR-PU 防御后的 t-SNE 图。若后门信息被有效清除,后门图像(黑色簇)将融入相应的干净图像簇中。图 5 显示,经过防御后,后门模型的 t-SNE 图中被污染的图像(黑色簇)从独立簇分散到对应的干净图像簇,证明了所提防御方法 DBR-PU 的有效性。

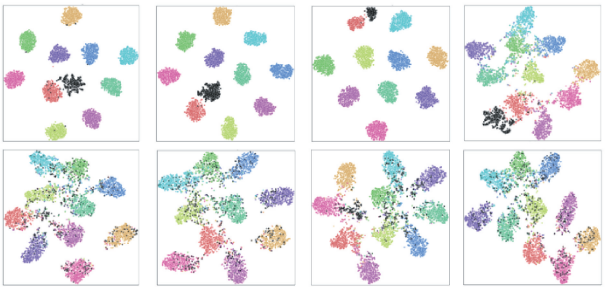

图 5　DBR-PU 在 CIFAR10 上的防御结果 t-SNE 可视化

Figure 5　t-SNE visualization of DBR-PU defense results on CIFAR10

3　结论

本文探索了在干净训练数据缺乏的情况下消除模型潜在的后门风险这一具有挑战性的场景,引入了一种新的基于剪枝和对抗性后门遗忘的无数据后门移除方法(DBR-PU)。该方法将基于分布熵的剪枝方法与无数据知识蒸馏相结合来克服缺乏干净训练数据这一条件限制,同时引入了对抗性后门遗忘模块来消除后门遗漏的风险。并在两个公开基准数据集中进行了大量的实验。实验结果显示,在缺乏原始训练数据的场景中,所提方法在模型准确率方面表现较好,在降低后门攻击成功率方面表现最好。此外,该方法对模型的原始精度的影响微乎其微,这表明本文所提方法在后门防御方面具有很好的性能。

参考文献:

[1] 罗荣辉,袁航,钟发海,等. 基于卷积神经网络的道路拥堵识别研究[J]. 郑州大学学报(工学版), 2019,40(2):21-25.
LUO R H, YUAN H, ZHONG F H, et al. Traffic jam detection based on convolutional neural network [J]. Journal of Zhengzhou University (Engineering Science), 2019,40(2):21-25.

[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] 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.

[4] NGUYEN A, TRAN A. WaNet: imperceptible warping-based backdoor attack[EB/OL]. (2021-02-20)[2025-08-16]. https://doi.org/10.48550/arXiv.2102.10369.

[5] BARNI M, KALLAS K, TONDI B. A new backdoor attack in CNNS by training set corruption without label poisoning[C]//2019 IEEE International Conference on Image Processing (ICIP). Piscataway: IEEE, 2019: 101-105.

[6] TRAN B, LI J, MADRY A. Spectral signatures in backdoor attacks[EB/OL]. (2018-11-01)[2025-08-16]. https://doi.org/10.48550/arXiv.1811.00636.

[7] WU D X, WANG Y S. Adversarial neuron pruning purifies backdoored deep models[EB/OL]. (2021-10-27)[2025-08-16]. https://doi.org/10.48550/arXiv.2110.14430.

[8] ZENG Y, CHEN S, PARK W, et al. Adversarial unlearning of backdoors via implicit hypergradient[EB/OL]. (2021-10-07)[2025-08-16]. https://doi.org/10.48550/arXiv.2110.03735.

[9] ZHENG R K, TANG R J, LI J Z, et al. Pre-activation distributions expose backdoor neurons[J]. Advances in Neural Information Processing Systems, 2022, 35: 18667-18680.

[10] ZHANG X Y, ZHOU X Y, LIN M X, et al. ShuffleNet: an extremely efficient convolutional neural network for mobile devices[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 6848-6856.

[11] CAI R, ZHANG Z Y, CHEN T L, et al. Randomized channel shuffling: minimal-overhead backdoor attack detection without clean datasets[J]. Advances in Neural Information Processing Systems, 2022, 35: 33876-33889.

[12] CHEN H T, WANG Y H, XU C, et al. Data-free learning of student networks[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway:IEEE, 2019: 3514-3522.

[13] FANG G F, SONG J, SHEN C C, et al. Data-free adversarial distillation[EB/OL]. (2019-12-23)[2025-08-16]. https://arxiv.org/abs/1912.11006.

[14] SHI L C, JIAO Y Y, LU B L. Differential entropy feature for EEG-based vigilance estimation[C]//2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Piscataway: IEEE, 2013: 6627-6630.

[15] CHEN X Y, LIU C, LI B, et al. Targeted backdoor attacks on deep learning systems using data poisoning[EB/OL]. (2017-12-15)[2025-08-16]. https://arxiv.org/abs/1712.05526.

[16] NGUYEN T A, TRAN A. Input-aware dynamic backdoor attack[J]. Advances in Neural Information Processing Systems, 2020, 33: 3454-3464.

[17] WANG Z T, ZHAI J, MA S Q. BppAttack: stealthy and efficient trojan attacks against deep neural networks via image quantization and contrastive adversarial learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 15074-15084.

[18] LIU K, DOLAN-GAVITT B, GARG S. Fine-pruning: defending against backdooring attacks on deep neural networks[C]//Research in Attacks, Intrusions, and Defenses. Cham: Springer, 2018: 273-294.

[19] WU B Y, CHEN H R, ZHANG M D, et al. Backdoorbench: a comprehensive benchmark of backdoor learning[J]. Advances in Neural Information Processing Systems, 2022, 35: 10546-10559.

[20] HEUSEL M, RAMSAUER H, UNTERTHINER T, et al. GANs trained by a two time-scale update rule converge to a local Nash equilibrium[EB/OL]. (2017-06-26)[2025-08-16]. https://arxiv.org/abs/1706.08500.

[21] VAN DER MAATEN L, HINTON G. Visualizing data using *t*-SNE[J]. Journal of Machine Learning Research, 2008, 9(11): 2579-2605.

Backdoor Removal Method for Deep Neural Networks Based on Pruning and Backdoor Unlearning

LI Xuexiang[1], GAO Yafei[1], XIA Huili[2], WANG Chao[1], LIU Minglin[1]

(1. School of Cyber Science and Engineering, Zhengzhou University, Zhengzhou 450002, China; 2. Henan Multimodal Perception and Intelligent Interaction Technology Engineering Research Center, Zhengzhou University of Economics and Business, Zhengzhou 451191, China)

Abstract: Backdoor attacks pose a serious threat to the security of deep neural networks. Most existing backdoor defense methods relied on partial original training data to remove backdoor from models. However, in real-world scenarios where these data access was limited, these methods performed poorly in eliminating backdoor and often significantly impact the model's original accuracy. To address these issues, in this study proposes a data-free backdoor removal method was proposed based on pruning and backdoor unlearning (DBR-PU). Specifically, the proposed method first analyzed the pre-activation distribution differences of model neurons on a synthetic dataset to identify suspicious neurons. Then, it reduced the impact of backdoor by pruning these suspicious neurons. Finally, an adversarial backdoor unlearning strategy was employed to further eliminate the model's internal response to any residual backdoor information. Extensive experiments on the CIFAR10 and GTSRB datasets against six mainstream backdoor attack methods demonstrated that, under data access constraints, the proposed method achieved a minimal accuracy gap compared to the best baseline defense methods and performed the best in reducing attack success rates.

Keywords: deep neural network; backdoor attack; backdoor defense; pre-activation distribution; adversarial backdoor unlearning