DOI:10.11918/202507065

基于反向遗忘的后门毒化样本检测

闫雷鸣,尤剑飞

(南京信息工程大学 计算机学院、网络空间安全学院,南京 210044)

摘 要:为提升模型性能,深度神经网络常需引入不可信数据集,导致易受数据投毒后门攻击。传统检测方法依赖识别毒化与良性样本的特征差异,但当攻击者优化触发器以模糊此边界时,其效果受限。针对此问题,本文提出反向遗忘(reverse forgeting, RFgt)检测方法,利用后门攻击中“毒化样本占比低”的特性,采用逆向优化策略:强制中毒模型快速遗忘占多数的良性样本特征,同时保留并强化对可疑样本的学习,以巩固其毒化特征,显著放大两类样本的特征差异,最终通过样本预测熵值判定是否为毒化样本。研究表明:RFgt 在 CIFAR-10 和 GTSRB 数据集上能够有效检测多种后门攻击下的毒化样本,同时对良性样本保持较低的误检率;在 Tiny ImageNet 数据集上的检测结果证明本方法具备良好的泛化能力。针对4种经典的数据投毒攻击,本方法平均检测真阳率达到99.28%,假阳率仅为0.06%,其综合性能优于现有防御方法。

关键词:后门攻击;数据毒化;样本检测;遗忘学习;预测熵

中图分类号:TP391 **文献标志码**:A **文章编号**:0367-6234(2026)05-0116-10

Backdoor poisoned sample detection via reverse forgetting

YAN Leiming, YOU Jianfei

(Nanjing University of Information Science and Technology, School of Computer Science, School of Cyber Science and Engineering, Nanjing 210044, China)

Abstract: To enhance model performance, Deep Neural Networks are frequently trained on untrusted datasets, rendering them vulnerable to data poisoning backdoor attacks. Conventional detection methods rely on identifying feature discrepancies between poisoned and benign samples. However, their effectiveness diminishes when attackers optimize trigger generation to obscure this boundary. To address this issue, this paper proposes a novel detection method named reverse forgeting (RFgt). The method exploits the characteristic of backdoor attacks, where the proportion of poisoned samples is low, and employs a reverse optimization strategy. Instead of forcing a poisoned model to forget backdoor features, RFgt compels it to rapidly forget the features of the majority class (benign samples), while simultaneously retaining and reinforcing the learning of suspicious samples to consolidate their poisoned features. This approach significantly amplifies the feature disparity between the two sample types. Ultimately, the prediction entropy of the samples is used to determine whether they are poisoned or benign. Experimental results demonstrate that RFgt effectively detects poisoned samples under various backdoor attacks on the CIFAR-10 and GTSRB datasets, while maintaining a low false positive rate. Furthermore, this method demonstrates strong generalization capability, as shown by its performance on the Tiny ImageNet dataset. Specifically, against four classic data poisoning attacks, RFgt achieves an average True Positive Rate (TPR) of 99.28% and a False Positive Rate (FPR) of only 0.06%, outperforming existing defense methods in overall performance.

Keywords: backdoor attack; data poisoning; sample detection; forgetting learning; predictive entropy

在深度学习模型的训练中,为了追求突出的性能,从网络公开渠道获取大规模数据集以扩充训练数据已成为常态。然而,这一开放式的做法将模型置于数据投毒后门攻击的直接威胁之下。攻击者通过在训练数据中注入少量精心设计的毒化样本,可以在模型中植入一个后门。该模型在处理正常输入时表现优异,不易察觉异常,但一旦输入中包含特定的触发器(Trigger),其预测结果便会被强制导向攻击者预设的目标标签。因此,如何有效地检测和剔除这些来自不可信来源的毒化样本,是至关重要的。

收稿日期:2025-07-26;**录用日期**:2025-09-28;**网络首发日期**:2025-10-22
网络首发地址:https://link.cnki.net/urlid/23.1235.T.20251021.1508.004
基金项目:国家自然科学基金(62172292,42375147)
作者简介:闫雷鸣(1973—),男,副教授,硕士生导师
通信作者:闫雷鸣,yan_leiming@163.com

传统的后门防御，其根基在于利用毒化过程产生的可观测的特点。已有研究[1]揭示，由于触发器与目标标签之间存在强行建立的、独特的因果关系，使得毒化样本比依赖广泛语义特征的良性样本更容易被模型学习和记忆。此外，在施加特定干扰[2]时，被逻辑捷径固化的毒化样本与依赖正常特征的良性样本，其模型预测结果会产生显著的偏离。早期的检测方法正是依赖于识别上述这些静态的、可分离的特征差异。然而，当前的攻防博弈已经进入一个新的阶段，攻击者不再留下易于察觉的痕迹，而是转向设计与良性样本特征高度融合的隐蔽触发器。通过精心调整触发器的强度及其与图像内容的融合度，攻击者可以在极低中毒率下高效触发恶意行为[3-4]，这使得毒化样本与良性样本在特征空间上高度重叠，从而直接瓦解了传统检测方法赖以生存的特征可分离设定。

面对这一困境，后门防御研究的关注点开始从静态特征转向动态学习范式。尽管毒化样本的设计日趋复杂与隐蔽，但其与良性样本在模型的学习范式上仍存在根本差异：良性样本的学习依赖于数据驱动的、广泛的特征关联，其模型记忆具有可塑性；而毒化样本的学习则依赖于由触发器主导的、固化的逻辑捷径，其模型记忆具有高度的抵抗性。利用这一差异设计防御策略已成为一个重要的研究方向。如 Wang 等[5]尝试通过“前置遗忘”范式优化反后门学习方法，在训练中直接移除可疑后门样本以消除后门。尽管该方法取得了一定效果，但其本质仍是一种直接对抗策略——试图正面清除抵抗性强的后门，在面对高隐蔽性后门时效果有限。

为此，本文提出转变防御思路，从被动地寻找微弱差异转向主动地创造显著差异。尽管毒化样本与良性样本在特征上高度相似，但在模型的学习范式上存在根本区别：1）良性样本的学习依赖于数据驱动的、分布广泛的语义特征，其在模型中的记忆具有可塑性；2）毒化样本的学习则依赖于由触发器主导的、高度相关的逻辑捷径，其记忆在模型中是固化的、抵抗性强的。

基于这一思路，本文提出了一种反向遗忘（RFgt）的后门毒化样本检测方法。其核心思想不是与后门特征进行直接对抗，而是利用良性样本数量占优，且其特征具有可塑性的特点，采取逆向策略：不试图让模型遗忘难以去除的后门特征，而是通过施加熵增约束，强制模型快速遗忘数量占优的良性样本特征。由于良性特征的可塑性，其在遗忘压力下会迅速衰退，导致模型对其预测的不确定性（熵）显著增加。相反，毒化样本因其固化的触发器捷径，对此遗忘过程表现出强大的抵抗性，其预测熵值保持在较低水平。这一过程主动地、非对称地改变了模型对两类样本的记忆，从而将它们之间原本微弱的动态学习差异，放大为预测熵值的差异，可为后续的精准检测奠定基础。最终，防御者仅需根据样本的预测熵值，即可高效、准确地区分出毒化样本。

1　后门攻击与后门防御

1.1　模型后门攻击

攻击者可从数据、模型参数、模型结构等多个角度嵌入后门触发器。根据嵌入方式，后门攻击可分为数据投毒攻击和模型中毒攻击。本文方法主要针对毒化样本进行检测，因此下文重点介绍数据投毒类后门攻击的发展脉络。

数据投毒攻击的核心在于触发器的设计，其演进趋势主要体现在隐蔽性的不断提升和对防御策略的适应性规避。早期的攻击，如 Gu 等[6]提出了基础的后门攻击方案（Badnets），采用简单的像素块作为触发器，虽然有效，但视觉上易于察觉。为了提升隐蔽性，后续研究转向设计“不可见”或“不易察觉”的触发器。如 Li 等[7]引入透明度控制来混合触发器与图像（Blended）；Turner 等[8]和 Zhong 等[9]则通过微小的像素值扰动或对抗性扰动生成触发器；Zhu 等[10]利用信息隐藏技术嵌入触发器，使其在统计层面难以被检测。

为了应对后门防御技术的发展，攻击者进一步探索了更为自然的触发器模式，使其与良性样本的特征分布更为接近。如 Liu 等[3]采用物理世界的反射现象（Refool），而 Nguyen 等[4]利用平滑的图像扭曲场（Wanet），这些触发器在视觉上几乎不留痕迹。同时，为了规避基于单一样本异常的检测器，一些高级攻击被设计出来。Barmi 等[11]通过仅修改目标类别的样本，并使用正弦信号作为触发器，实现了干净标签攻击。Qi 等[12]将触发器分片嵌入不同图像，增加了检测难度；Mo 等[13]则通过在数据集中添加混淆样本，直接干扰防御方法的决策边界。

目前后门攻击正朝着高隐蔽性、特征融合以及防御规避的方向发展。这种趋势使得毒化样本与良性样本在原始特征空间（无论是像素空间还是频域空间）的差异变得十分微小，给依赖静态特征差异进行检测的传统防御方法带来了巨大挑战。

1.2　模型后门防御

传统的后门防御方法可大致分为 3 类：数据集后门检测、输入后门检测和模型后门检测。

数据集后门检测致力于在模型训练前或训练过程中净化数据集。这类方法普遍基于一个核心假

设:毒化样本在某个特征空间中会呈现出与良性样本不同的分布。如 Tran 等[14]和 Zeng 等[15]假设触发器会在频域留下痕迹,并利用光谱特征或高频伪影进行检测。Hayase 等[16]和 Pan 等[17]则通过分析样本在深度特征空间中的表征来识别异常。然而,这些方法共同的局限性在于,其依赖于一种可分离的、静态的特征差异。如 1.1 节所述,后门攻击通过精心设计触发器,能够有效模糊这种特征差异,从而规避检测。

输入后门检测发生在模型部署阶段,旨在识别携带触发器的恶意输入。这类方法通常采用扰动输入或分析模型响应来判断。如 Gao 等[18]通过叠加图像检测预测的随机性;Chou 等[19]通过复制可疑区域观察预测变化;Liu 等[20]分析模型对图像损坏响应的一致性。Chen 等[21]则试图通过重构输入来破坏触发器。这些方法的优势在于实时防御,但其主要挑战在于需要处理的样本是孤立的,缺乏全局数据分布信息,且可能引入额外的推理延迟。其旨在"拦截"攻击,而非"根除"源头(即净化数据集)。

模型后门检测旨在判断一个给定的模型是否已被植入后门。如 Wang 等[22]通过逆向工程还原触发器模式(neural cleanse);Zheng 等[23]则利用模型拓扑结构的差异进行检测。Liu 等[24]将黑白盒作为测试用例检测模型是否存在后门。这类方法能够对模型安全性进行宏观评估,但通常无法定位到具体的毒化样本,因此,对于需要净化数据集以重新训练一个干净模型的防御者而言,其作用有限。

现有后门防御工作大多依赖于在某个静态特征空间中识别毒化样本与良性样本的固有差异。无论是数据集净化、输入检测还是模型诊断,其成功都以"特征可分离"这一假设为前提。然而,高级后门攻击通过精心设计的触发器,使得毒化样本在特征层面与良性样本高度融合。当特征差异被攻击者主动抹除时,这些防御方法的性能便会显著下降。本研究通过引入反向遗忘机制,主动干预模型对不同样本的记忆和遗忘行为,从而将检测的依据从难以区分的"输入特征"转移到被显著放大的"模型响应差异"上。

2 基于反向遗忘的后门毒化样本检测方法

2.1 防御场景

攻击者通过数据毒化的方式,在训练数据中构造两个部分:良性样本(x,y)和毒化样本(x_p,y_t)。良性样本表示未受污染的输入及其正确标签,而毒化样本表示带有触发器 T 的输入 $x_p=x+T$ 及其目标标签 y_t。为了同时满足模型在正常任务上的性能和后门触发效果,攻击者的训练目标可以统一描述为

$$\arg\min_{\theta} L = E_{(x,y)}L(f_\theta(x),y) + E_{(x_p,y_t)}L(f_\theta(x_p),y_t) \tag{1}$$

式中:θ 为模型的参数;L 为分类损失函数,用于衡量模型预测$f_\theta(x)$与真实标签 y 之间的差距;E 为对样本的期望计算。

通过最小化上述损失函数,攻击者确保模型在干净数据上的分类性能不受影响,同时嵌入后门,使得模型在接收到触发器样本时输出攻击者指定的目标类别。

在许多实际应用场景中,防御者(如企业或研究机构)通常能够从自身业务中获得一小部分规模有限但来源可靠、可信的良性样本集D_c。然而,仅凭这些数据通常不足以训练出性能卓越的深度学习模型。为了弥补自有样本量的不足并提升模型泛化能力,防御者不得不从外部(如公共数据集、第三方数据提供商)获取大规模的未知数据集D_p进行补充。因此,本文将"持有少量可信样本D_c"作为防御起始点是符合实际情况的合理设定,核心挑战在于如何利用这部分少量可信数据,来净化大规模的、潜在被污染的外部数据集D_p。同时,防御者无法预知外部数据中是否存在毒化样本,也无法确定其比例 ε。因此,在缺乏具体攻击信息的条件下,防御方法需要通过有效的策略设计,识别毒化样本的异样特征,从而剔除数据集中的毒化样本,最后得到正常的数据集。

2.2 方法流程与核心机制

本文提出的基于反向遗忘的后门毒化样本检测方法(RFgt),是一种分阶段、逐步放大特征差异的检测策略。该策略的核心是通过一个动态且非对称的遗忘过程,将毒化样本与良性样本在学习范式上的内在差异,转化为一个易于度量的预测熵差异。反向遗忘的整体框架如图 1 所示,该流程被系统地划分为 3 个阶段。

2.2.1 初步遗忘与样本扩充

防御者仅持有一小部分可信的良性样本集D_c,以及一个大规模的、来源不可信的待检测数据集D_p,并可以利用中毒数据集D_p训练一个具备初步判别能力的初始模型M_b。若直接利用规模过小的D_c对初始中毒模型M_b施加反向遗忘,其产生的约束效应将不足以显著改变模型对D_p中良性样本的既有记忆,因而无法形成有效的熵值分化。为解决这一数据不平衡问题,本文设计了一种样本扩充机制。首先,利用D_c对M_b施加一个初步的反向遗忘训练,

得到初步遗忘模型 M'。此阶段的目标并非实现完全遗忘，而是作为一次初始扰动，其损失函数 L_e 旨在最大化D_c中良性样本的预测熵。

$$L_e = -H(x) = \sum_{i=1}^{C} p(y_i|x)\log p(y_i|x) \quad (2)$$

式中：$p(y_i|x)$表示模型对输入样本 x 的预测概率，C为类别数，$H(x)$为样本 x 的整体预测熵。

该损失函数的目标是通过最大化熵，迫使模型反向遗忘良性样本的特征，对良性样本做出更为不确定的预测。这一阶段利用少量良性样本D_c引导模型迅速忘记良性样本特征，得到初步遗忘模型 M'。少量良性样本D_c在初步遗忘模型 M'中的特征会因反向遗忘机制而迅速衰减，模型 M'的学习重点会由原来的良性样本特征转向毒化样本特征。通过该损失函数，模型 M'能够提升对良性样本预测的不确定性，同时确保毒化样本的特征不会被过度遗忘。

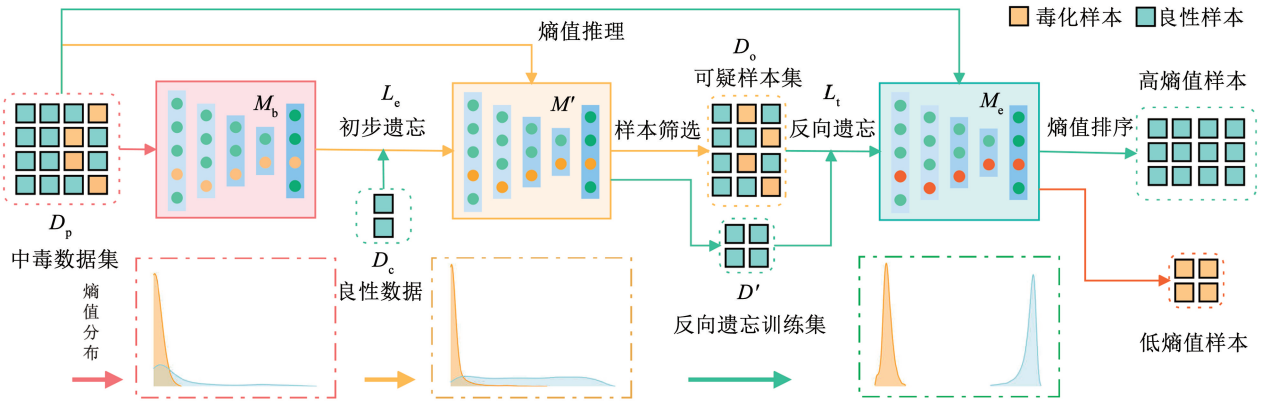


图 1　反向遗忘的整体框架图

Fig. 1　Overall framework of reverse forgetting

基于此现象，将D_p中所有样本输入 M'，筛选出预测熵最高的一部分样本。这些高熵样本可被视为经初步遗忘识别出的高置信度良性样本。将其与初始D_c合并，构成一个规模和代表性均得到增强的反向遗忘训练集 D'，剩余样本则构成可疑样本集D_o，

$$D_p = D' \cup D_o \quad (3)$$

2.2.2　深化反向遗忘

在获得规模充分的反向遗忘训练集 D'后，即可执行核心的深化反向遗忘操作，最终得到一个反向遗忘模型 M_e。实验中对 D'和D_o施加不同的训练目标，以最大化两类样本在模型响应上的最终差异。该目标的复合损失函数 L_t 定义为

$$\underset{\theta}{\operatorname{argmin}}\, L_t = \alpha E_{(x,y)} L(f_\theta(x), y) + \alpha E_{(x_p,y_t)} L(f_\theta(x_p), y_t) - \beta E_{(x_c,y)} H(f_\theta(x_c)) \quad (4)$$

式中：α 为交叉熵损失项的权重，β 为增熵约束项的权重，θ 为模型参数，$f_\theta(x)$为模型对样本 x 的推理输出，E 为对样本的期望计算，x_c 为少量良性样本集 D_c 中的样本。

该损失函数包含两个功能不同的组成部分。1）对D_o的标准交叉熵学习：促使模型继续在D_o上进行标准的分类学习。此设计的目的在于维持并强化毒化样本的低熵特性，保证其“触发器 - 目标标签”的映射关系不被破坏。2）对 D'的熵最大化约束：第二项则对 D'中的样本施加强力的熵最大化约束，驱动模型对这些高置信度良性样本的预测趋向于均匀分布，使其预测熵达到理论最大值（ln10）。

在反向遗忘的训练过程中，增熵约束项的目的是让良性样本的预测熵尽快达到最高。在 CIFAR-10 数据集上，如果一个样本没有被学习，则该样本可能被分到 10 个类别中的任意 1 个类别，即分到每个类别的概率均为 1/10。根据离散随机变量的信息熵的定义，单个样本的最大熵值为 ln10，则增熵约束就是为了让每个良性样本的预测熵值尽可能达到 ln10。

通过对一部分正常数据施加遗忘约束，同时，对另一部分异常数据维持标准学习，模型中两类样本的学习动态被导向截然不同的方向。该过程将它们之间原本微弱的内在差异，显著地放大为模型M_e输出端一个清晰、可分离的熵值分布间距。

2.2.3　毒化样本检测

根据第 2 阶段训练过程中模型M_e所学到的特性，良性样本具有较高的预测熵值，而毒化样本则表现为较低的预测熵值。为了有效区分良性样本与毒化样本，设定 1 个阈值 T。具体判断规则如下：当样

本的预测熵值 H 大于等于阈值 T 时,样本被认为是良性样本;当样本的预测熵值小于阈值 T 时,样本则被归类为毒化样本。

为了合理地确定阈值 T,本文采用一种基于聚类的双中心点划分策略:首先通过 k-means 聚类算法将预测熵值数据 $E=\{e_1,e_2,\cdots,e_n\}$ 强制分为两类 C_1 和 C_2,对应两组数据中心点 μ_1 和 μ_2(假设 $\mu_1>\mu_2$),其中 e 为先前得到的单个样本的预测熵值。聚类目标函数为

$$\min_{C_1,C_2}\sum_{e\in C_1}(e-\mu_1)^2+\sum_{e\in C_2}(e-\mu_2)^2 \tag{5}$$

式中:$\mu_1=\frac{1}{|C_1|}\sum_{e\in C_1}e$,$\mu_2=\frac{1}{|C_2|}\sum_{e\in C_2}e$。

此时阈值 T 可定义为两中心点的加权中间值,表达式为

$$T=\frac{\mu_1+\mu_2}{2}+\lambda\cdot\left(\frac{\mu_1-\mu_2}{2}\right) \tag{6}$$

式中 λ 为调整系数。通过这种方式,模型 M_e 可以有效地区分良性样本与毒化样本,实现对毒化样本的检测。

3 实验结果与分析

3.1 实验设置

本研究使用了两个广泛应用于图像分类任务的数据集:CIFAR-10 和 GTSRB 数据集。CIFAR-10 数据集包含 10 个类别的彩色图像,每张图像的尺寸为 32×32 像素,数据集共包含 60 000 张图像,其中 50 000 张用于训练,10 000 张用于测试。GTSRB 数据集是交通标志识别任务的数据集,包含 43 种不同类型的交通标志,共计 39 209 张训练图片和 12 630 张测试图片。为了进一步验证方法的泛化能力,本研究还引入了 Tiny ImageNet 数据集。该数据集是 ImageNet 的一个子集,包含 200 个类别,有 100 000 张训练集图像和 10 000 张验证集图像,图像尺寸为 64×64 像素。

本文选择 6 种先进的数据投毒后门攻击方法,包括 Badnets、Blend、SIG、CL、Wanet 和 Refool。攻击的目标标签默认为 0。除了干净标签攻击,每种后门攻击的样本投毒率均设定为 5%。

本文的主要实验分为两部分:1)通过少量良性样本 D_c 进行初步遗忘筛选高熵样本扩充数据;2)使用扩展后的数据 D' 进行反向遗忘。初始模型训练使用 ResNet-18 进行训练,在第 1 部分中,初始的良性样本数量设为数据集样本数的 2% ~5%,通过 5 轮的熵约束后筛选熵值最高的 10% ~20% 的样本进行扩充。在进行反向遗忘时,交叉熵损失项的权重参数 α 为 1,熵约束项的权重参数 β 为 3,初始学习率设为 0.001,优化器选择 Adam 优化器,训练轮数设定为 50 轮。

3.2 对比实验

为确保对 RFgt 方法性能进行准确的评估,本文选取 ASD[25]、DE[26] 和 BDE[27] 三种先进的防御方法作为对比基准。该选择旨在将 RFgt 置于当前主流且多样化的防御策略中进行考量,具体而言:ASD 代表了基于半监督学习的防御思路,通过识别良性样本与可疑样本在表征空间中的分布差异来净化数据,其有效性依赖于特征的可分离性假设;DE 则利用扩散模型,通过对样本进行去噪重构来主动擦除潜在的后门触发器,是生成式净化方法的典型代表;BDE 是一种基于模型微调的防御方法。重要的是,该方法与本文工作共享完全相同的防御前提,即利用一小部分可信的干净数据进行防御。因此,将 BDE 作为基准,能够为 RFgt 的有效性提供一个直接的参照,而与 ASD 和 DE 的比较,则能够检验 RFgt 在面对不同防御方案时的相对优势。

在得到 RFgt 方法检测样本后的数据集后,本研究使用该数据集在 ResNet-18 网络上训练一个新的分类模型,所训练出的模型的准确率(ACC)和攻击成功率(ASR)如表 1 所示。由表 1 结果可以看出,RFgt 方法能够有效防御所有 6 种评估的后门攻击,在 CIFAR-10 数据集上平均达到 93.4% 的模型准确率和 0.7% 的攻击成功率,在 GTSRB 数据集上平均达到 96.6% 的模型准确率和 0.7% 的攻击成功率。这一结果证明,RFgt 能够在有效剔除毒化样本的同时,保留良性样本的数据价值。

表 1 中还展示了其他 3 种防御方法的模型表现。ASD 方法采用了半监督学习,通过从大量未标记数据中筛选样本来防止后门投毒,表现出较低的 ASR。DE 方法利用扩散模型消除毒化样本的特征,并恢复良性样本的特征,其 ACC 略低于本文方法,但 ASR 表现优异。BDE 方法同样采用少量干净数据微调训练好的后门模型,但在 CIFAR-10 数据集上有 6.1% 的 ASR,在测试的所有防御方法中最高。这是因为 BDE 方法在微调后门模型的过程中,需要修改良性样本标签来完成训练,容易干扰对特征高度相似的毒化样本识别判断。而 RFgt 方法全程无需改动样本标签,依靠熵值变化主动拉大两类样本差异,能够有效规避后门毒化带来的负面影响,检测稳定性更强。

表 1　多种防御方法对抗 6 种后门攻击时的防御表现

Tab. 1　Defense performance of our method and other approaches against six backdoor attacks

数据集	攻击方式	无防御		BDE[27]		ASD[25]		DE[26]		RFgt(本文方法)	
		ACC/%	ASR/%	ACC/%	ASR/%	ACC/%	ASR/%	ACC/%	ASR/%	ACC/%	ASR/%
CIFAR-10	Badnets[6]	94.1	100.0	93.1	0.1	93.3	1.5	92.9	0.7	**94.0**	**0**
	Blended[7]	93.9	98.9	93.0	8.7	93.1	1.8	93.2	1.3	**94.1**	**0.5**
	SIG[11]	93.6	95.3	92.3	6.3	92.7	1.3	92.5	2.1	**93.8**	**0.4**
	Wanet[4]	93.1	98.7	91.9	5.8	92.6	2.7	92.3	**0.2**	**93.0**	0.3
	Refool[3]	93.7	99.8	92.3	3.2	93.4	0.9	92.7	**0.8**	**93.8**	1.3
	A-Patch[12]	92.9	97.6	92.4	12.3	93.1	5.7	91.6	1.9	**91.9**	**0.8**
	SSDT[13]	93.4	98.6	91.6	6.5	92.5	3.6	92.3	2.2	**93.1**	**1.5**
	Average	93.5	98.4	92.3	6.1	92.9	2.5	92.5	1.3	**93.4**	**0.7**
GTSRB	Badnets	97.3	100.0	96.3	0.1	96.1	0.4	96.5	0.7	**97.4**	**0**
	Blended	97.1	99.8	95.8	4.2	96.3	0.5	96.3	1.2	**97.0**	**0.1**
	SIG	96.5	97.3	95.2	8.9	95.9	1.3	96.2	1.9	**96.3**	**0.8**
	Wanet	97.1	99.9	94.7	6.2	**96.3**	0.6	96.0	**0.2**	96.2	1.4
	Refool	97.2	99.3	95.3	6.7	96.4	0.8	96.3	1.4	**96.8**	0.4
	A-Patch	96.8	95.6	96.4	15.6	96.1	4.2	95.3	2.4	**95.7**	**1.4**
	SSDT	97.1	98.3	95.5	6.2	96.3	4.4	96.1	1.6	**96.8**	**0.9**
	Average	97	98.6	95.6	6.8	96.2	1.7	96.1	1.3	**96.6**	**0.7**

表 2 展示了在中毒率为 5% 的 CIFAR-10 数据集上，本文提出的 RFgt 方法与 BDE、ASD、DE 三种主流先进防御方法的样本检测效果对比，通过检测成功率(TPR)和误检率(FPR)两项关键指标，量化评估各方法的检测精度与误判风险。从表 2 数据可知，RFgt 方法表现出最优的综合检测性能，其平均检测成功率达到 99.28%，同时平均误检率仅为 0.06%，在所有对比方法中同时实现了最高的 TPR 和最低的 FPR。

RFgt 方法能够同时实现高 TPR 与低 FPR，其关键在于反向遗忘机制所产生的双重效应。一方面，该机制对高置信度良性样本施加全局性的熵最大化约束，使其预测分布一致地向高熵状态转变。该过程不同于传统的离群点检测，其从机制上避免了将特征位于分布边缘的良性样本误判为异常，因而导致了极低的 FPR。另一方面，在强制遗忘良性特征的同时，对包含毒化样本的可疑集维持了标准的交叉熵学习。这不仅保留了毒化样本因逻辑捷径产生的低熵特性，更在该反向遗忘过程的约束下，使其低熵状态相对更加稳固和易于区分，从而实现了高 TPR。

表 2　在中毒率 5% 的 CIFAR-10 数据集上的样本检测效果

Tab. 2　Sample detection performance on the CIFAR-10 dataset with 5% poisoning rate

攻击方式	BDE		ASD		DE		RFgt(本文方法)	
	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
Badnets	94.76	0.68	99.67	4.62	99.40	0.89	**100.00**	**0**
Blended	96.64	1.48	99.44	4.39	99.18	1.15	**99.98**	**0**
SIG	90.60	0.49	93.54	7.16	94.48	0.92	**99.83**	**0**
Wanet	88.35	8.94	92.43	5.61	**99.40**	1.20	97.32	**0.24**
Average	92.59	2.90	96.27	5.45	98.12	1.04	**99.28**	**0.06**

相比之下，ASD 方法虽然 TPR 较高，但其较高的 FPR 表明其半监督筛选策略较为激进，牺牲了部分良性样本。DE 方法虽然整体表现良好，但其 FPR 略高于 RFgt，这可能是因为其扩散净化过程在某些情况下会过度平滑图像特征，使得一些原本特征独特的良性样本的分布被改变，从而增加了被误判的风险。

为了验证 RFgt 方法在更复杂场景下的有效性，本文在类别更多、图像更复杂的 Tiny ImageNet 数据集上开展了泛化性验证实验，选取 4 种代表性后门

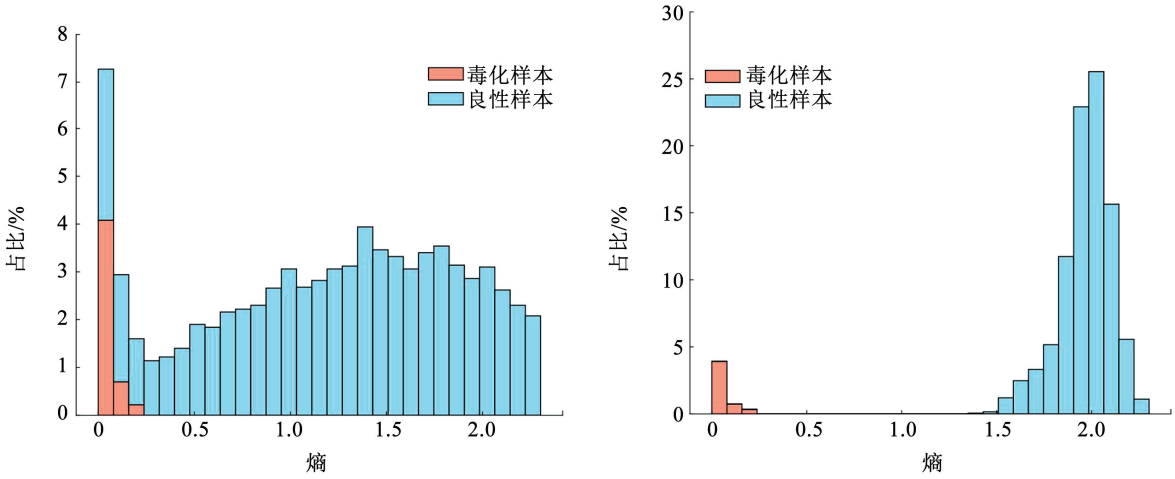
攻击方法进行评估。表 3 呈现了 RFgt 方法在该数据集上的防御表现,核心通过净化后模型的准确率(ACC)和攻击成功率(ASR)两项关键指标,量化衡量方法在大规模复杂数据集上的实际防御效果。实验结果表明,尽管 Tiny ImageNet 的数据复杂性和类别数量较之前的数据集大幅提升,但 RFgt 方法依然保持了卓越的防御性能,其净化后模型的平均准确率达 60.49%,攻击成功率则被抑制在 0.81% 的极低水平。这一结果表明,RFgt 方法不局限于小尺寸、少类别的简单任务,在更具挑战性的大规模数据集上同样具有适用性和鲁棒性。

RFgt 方法主要采用熵约束来遗忘良性样本的特征,为了验证该熵约束方案的可靠性,本文将 BDE 方法中通过更改样本标签混淆正常样本特征的遗忘方案,与本文提出的熵约束方案进行对比分析。图 2 展示了两种遗忘方案在 CIFAR-10 数据集与 Blended 中毒攻击下的样本熵值分布情况,用于直观对比不同遗忘策略对毒化样本与良性样本的特征分离效果。图中横坐标为不同样本的熵值,纵坐标为对应熵值区间的样本在完整数据集中的百分比占比。在防御 Blended 中毒攻击时,BDE 的标签修改方案虽然能够提高多数良性样本的熵值,但仍无法完全分离毒化样本与良性样本的熵值分布;而本文提出的熵约束策略能够使两类样本的熵值分布形成清晰的边界,为后续毒化样本的检测提供了明确的判别依据。

表 3 RFgt 方法在 Tiny ImageNet 数据集上的防御表现

Tab. 3 Defense performance on the Tiny ImageNet dataset

攻击方式	无防御		RFgt(本文方法)			
	ACC/%	ASR/%	TPR/%	FPR/%	ACC/%	ASR/%
Badnets	62.64	100.00	100.00	0.00	61.27	0
Blended	61.84	95.21	99.63	0.07	60.85	0.24
Wanet	61.73	95.93	97.32	0.85	60.64	1.84
Refool	62.53	96.23	98.13	0.41	59.19	1.16
Average	62.19	96.84	98.77	0.58	60.49	0.81



(a) 更改样本标签 (b) 熵值约束

图 2 不同遗忘方法在 CIFAR-10 数据集上的样本熵值分布

Fig. 2 Entropy distribution of samples on the CIFAR-10 dataset under different forgetting methods

3.3 消融实验

RFgt 方法基于少量的良性样本对中毒模型进行反向遗忘。为了探讨干净数据的数量对实验结果的影响,本文评估了 RFgt、BDE 和 ASD 方法在不同数量的良性样本数据上针对 CIFAR-10 的 Blended 攻击的表现。3 种方法的 TPR 和 FPR 变化如图 3 所示,随着良性样本数据量的减小,RFgt 方法的 TPR 仍然保持着较高的数值,仅在初始良性样本极少时下降至约 93%。相比之下,在相同条件下,ASD 方法的 FPR 较高,而 BDE 方法的 TPR 则表现不佳。

本文的防御方法要通过直接熵增约束扩充遗忘样本,扩充的样本数量会影响到最终反向遗忘的效果。为了研究扩充样本数量对实验效果的影响,设计了在 CIFAR-10 数据集受到不同中毒率的 blended 攻击时,扩充不同比例 ε 的样本对模型 ACC 和 ASR 的影响。从表 4 可以看出,当中毒率为 5% 且扩充 20% 的样本时的防御效果最好,随着中毒率和扩充样本数的增加,模型的 ASR 逐渐升高,这是由于扩充样本时混入了一小部分的毒化样本。同时,扩充样本比例 ε 过小会导致在遗忘阶段对良性样本的遗忘效果变差,从而导致模型准确率(ACC)较低。但在实际场景中,攻击者通常不会提供中毒率过高的数据集($\geqslant$20%),大量的毒化样本很容易被发现。

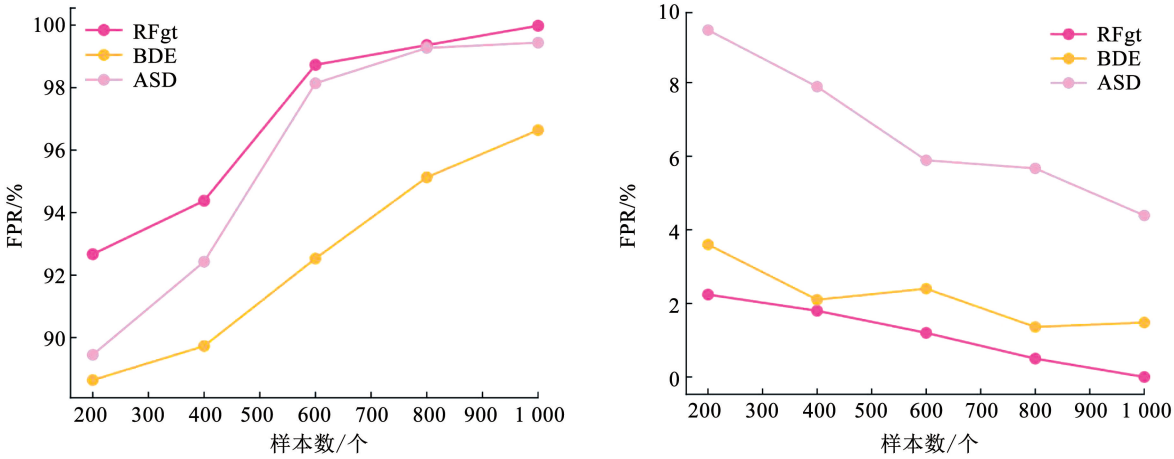


图 3　不同初始样本数量下 3 种防御方法的检测成功率和误检率的变化

Fig. 3　Changes in TPR and FPR of three defense methods under varying initial benign sample quantities

表 4　扩充不同比例 ε 的样本在 CIFAR10 数据集上的防御效果

Tab. 4　Analysis of RFgt method's defense effect with expanded sample ratios of ε on the CIFAR-10 dataset (Blended)

中毒率/%	ε=5%		ε=10%		ε=20%		ε=30%	
	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
5	91.51	0.64	93.78	0.48	94.12	0.51	94.03	1.31
10	92.23	0.63	93.89	0.35	93.72	0.83	93.89	1.43
20	91.74	0.83	93.21	0.73	93.53	1.25	93.77	2.57

由表 5 可见，RFgt 方法在处理低中毒率的 3 种后门攻击时均表现出良好的检测性能，检测样本的假阳率均为 0%，且在 1% 中毒率的情况下，仍有较高的真阳率。这是因为 RFgt 方法会使良性样本的整体特征被模型遗忘，检测的假阳率基本为 0。在真实世界的攻击中，攻击者为了增强隐蔽性，常采用极低的毒化率，本方法在 1% 的极低毒化率下依然能实现检测(FPR 为 0%)，充分证明了其在实际防御场景中的鲁棒性和应用价值。

表 5　RFgt 方法在 CIFAR10 数据集上对低中毒率攻击的检测效果

Tab. 5　Detection performance of the RFgt method against low-poisoning-rate attacks on the CIFAR-10 dataset

攻击方式	中毒率=1%		中毒率=3%		中毒率=5%	
	TPR	FPR	TPR	FPR	TPR	FPR
Badnets	100.00	0	100.00	0	100.00	0
Blended	99.37	0	99.82	0	99.98	0
SIG	98.54	0	99.21	0	99.83	0

超参数 α 与 β 分别决定了 RFgt 方法中保留毒化样本特征和遗忘良性样本特征的强度，其比值的合理取值是保证方法检测性能的关键。为了探究 β/α 比值对检测效果的影响，本文设计了超参数敏感性实验，实验结果如表 6 所示。由表 6 可以看到：随着 β/α 比值的增大，RFgt 的 TPR 呈下降趋势；反之，当该比值较小时，FPR 则会上升。其原因在于：若 β 权重过高，会过度遗忘特征，甚至影响到毒化样本的低熵特性，从而降低 TPR；若 β 权重过低，则遗忘效果不足，导致良性样本的熵值提升不充分，增加了其被误判的风险(即更高的 FPR)。

表 6　RFgt 方法在 CIFAR10 数据集上对不同攻击的检测效果

Tab. 6　Detection performance of the RFgt method against different attacks on the CIFAR-10 dataset

攻击方式	β/α=1		β/α=3		β/α=5	
	TPR	FPR	TPR	FPR	TPR	FPR
Badnets	100.00	0.12	100.00	0	98.53	0
Blended	100.00	0.78	99.98	0	99.12	0
Wanet	98.73	1.83	97.32	0.24	95.44	0

3.4　局限性与未来工作

本文提出的 RFgt 方法在检测数据投毒攻击方面表现出色，但其适用范围与机制设计亦存在一定的局限性，这为未来的研究指明了方向。

3.4.1　适用攻击类型的局限性

本文的核心工作是识别和剔除训练集中的毒化样本，因此主要适用于数据投毒攻击场景。对于模型投毒攻击，即攻击者直接修改已训练完成的模型参数来植入后门，由于其训练数据本身可能是完全良性的，数据集中不存在可供检测的“毒化样本”。RFgt 方法通过分析样本在遗忘过程中的动态变化来区分异常，这一机制依赖于对数据样本的直接操作，因此无法直接应用于检测一个给定的、静态的、已中毒的模型。未来的工作可以探索如何将“反向遗忘”的思想迁移至模型层面，如通过分析不同神

经元或网络层对良性数据特征的遗忘速率,来识别可能与后门功能相关的异常结构。

3.4.2 对初始良性样本的依赖

RFgt 方法依赖于一小部分可信的良性样本集 D_c。图 3 结果表明,本文方法在良性样本数量极少(如数据集的 2%)时依然有效,但其性能的下限仍受限于这部分初始数据的数量和代表性。在完全无监督的场景,即防御者没有任何可信数据的情况下,如何有效初始化反向遗忘过程,是一个值得探索的方向。

3.4.3 计算开销分析

本文所提 RFgt 方法的计算开销主要来源于其多阶段的训练流程,具体包括:初始后门模型的训练、用于样本扩充的初步反向遗忘,以及核心的深化反向遗忘。为客观评估其运行效率,本节将 RFgt 与 BDE、ASD、DE 三种对比方法在相同硬件环境下的总时间消耗进行了记录与分析,具体数据如表 7 所示。

表 7 多种防御方法在不同数据集上的时间消耗

Tab. 7 Time consumption (in hours) of various defense methods across different datasets h

数据集	防御方法			
	BDE	ASD	DE	RFgt
CIFAR-10	0.93	0.64	2.31	0.75
GTSRB	0.91	0.58	2.12	0.71

从表 7 的数据可以看出,RFgt 方法的计算效率具有显著竞争力。相较于 DE 方法,RFgt 的耗时大幅减少,这主要因为 DE 依赖于计算密集型的扩散模型进行多步样本去噪与重构。同时,RFgt 也比基于模型微调的 BDE 方法更高效。尽管 ASD 方法的耗时略低于 RFgt,但这种微弱的时间优势是以牺牲大量防御性能为代价的。结合表 2 的检测结果,ASD 在 CIFAR-10 上的平均 FPR 高达 5.45%,远高于 RFgt 的 0.06%。这表明,RFgt 通过增加有限的计算投入(约 0.11 h),换取了检测精度的提升。

尽管 RFgt 方法的总体时间成本在可接受范围内,且优于部分主流方法,但其多阶段的流程设计相较于单次检测方法,在计算步骤上更为复杂。虽然这一设计是其高性能的保证,但在部署效率要求极高的场景下,仍有优化的空间。

未来研究可侧重如何简化遗忘流程或采用更高效的训练策略,如设计端到端的遗忘框架,以进一步降低时间成本。

4 结 语

针对高级后门攻击能够将毒化样本伪装成为与良性样本高度相似,从而导致传统检测方法失效这一难题,本文提出了一种基于反向遗忘的后门样本检测方法(RFgt)。其核心创新在于,利用了良性样本记忆具有可塑性、而毒化样本记忆具有固化性的根本差异。RFgt 通过施加熵增约束,强制模型遗忘数量占优的良性特征,而非直接对抗隐蔽的后门特征。这一逆向策略能够将两类样本间原本难以察觉的学习动态差异,高效地放大为清晰可辨的预测熵差异。在 CIFAR-10 和 GTSRB 数据集上的实验表明,RFgt 方法能有效防御多种后门攻击。净化后模型的平均准确率分别达到 93.4% 和 96.6%,攻击成功率则均被抑制在 0.7% 的极低水平。在样本检测层面,其平均真阳率(TPR)高达 99.28%,而假阳率(FPR)仅为 0.06%。此外,在类别更多、图像更复杂的 Tiny ImageNet 数据集上的成功验证,也证明了本方法具备良好的泛化能力,可为深度学习模型抵御数据投毒后门攻击提供一种可靠的防御方案。

参考文献

[1] ZHANG Zaixi, LIU Qi, WANG Zhicai, et al. Backdoor defense via deconfounded representation learning[C]// Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, BC, Canada: IEEE, 2023: 12228. DOI: 10.1109/CVPR52729.2023.01177

[2] CHEN Weixin, WU Baoyuan, WANG Haoqian. Effective backdoor defense by exploiting sensitivity of poisoned samples[J]. Advances in Neural Information Processing Systems, 2022, 35: 9727. DOI: 10.48550/arXiv.2206.02704

[3] LIU Yunfei, MA Xingjun, BAILEY J, et al. Reflection backdoor: A natural backdoor attack on deep neural networks[C]// Proceedings of the European Conference on Computer Vision (ECCV 2020). Berlin, Germany: Springer, 2020: 182

[4] NGUYEN A, TRAN A. Wanet-imperceptible warping-based backdoor attack [C]// Proceedings of the International Conference on Learning Representations (ICLR). Virtual: OpenReview.net, 2021: 6667

[5] 王晗旭, 李欣, 许文韬, 等. 遗忘学习前置的反后门学习方法研究[J]. 计算机工程与应用, 2024, 60(19): 259
WANG Hanxu, LI Xin, XU Wentao, et al. Anti-backdoor learning method based on preposed unlearning[J]. Journal of Computer Engineering & Applications, 2024, 60(19): 259

[6] GU Tianyu, LIU Kang, DOLAN-GAVITT B, et al. BadNets: Evaluating backdooring attacks on deep neural networks[J]. IEEE Access, 2019, 7: 47230. DOI:10.1109/ACCESS.2019.2909068

[7] LI Shaofeng, XUE Minhui, ZHAO Benjamin, et al. Invisible backdoor attacks on deep neural networks via steganography and regularization[J]. IEEE Transactions on Dependable and Secure Computing, 2020, 18(5): 2088. DOI: 10.1109/TDSC.2020.3021407

[8] TURNER A M. Exploring the landscape of backdoor attacks on deep

neural network models[D]. Cambridge, MA, USA: Massachusetts Institute of Technology, 2019

[9]ZHONG Haoti, LIAO Cong, SQUICCIARINI A C, et al. Backdoor embedding in convolutional neural network models via invisible perturbation[C]// Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy (CODASPY). New York, NY, USA: ACM, 2020: 97. DOI:10.1145/3374664.3375751

[10]朱淑雯, 罗戈, 韦平, 等. 隐蔽图像后门攻击[J]. 中国图象图形学报, 2023, 28(3): 864. DOI: 10.11834/jig.220550
ZHU Shuwen, LUO Ge, WEI Ping, et al. Covert image backdoor attacks[J]. Journal of Image and Graphics, 2023, 28(3): 864. DOI: 10.11834/jig.220550

[11]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, NJ, USA: IEEE, 2019: 101. DOI: 10.1109/ICIP.2019.8802997

[12]QI Xiangyu, XIE Tinghao, LI Yiming, et al. Revisiting the assumption of latent separability for backdoor defenses[C]// Proceedings of the International Conference on Learning Representations (ICLR). Virtual: OpenReview.net, 2023: 19637

[13]MO Xiaoxing, ZHANG Yechao, ZHANG L Y, et al. Robust backdoor detection for deep learning via topological evolution dynamics[C]// 2024 IEEE Symposium on Security and Privacy (SP). Piscataway, NJ, USA: IEEE, 2024: 2048. DOI: 10.1109/SP54263.2024.00174

[14]TRAN B, LI J, MADRY A. Spectral signatures in backdoor attacks[C]//Advances in Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates, Inc., 2018: 801

[15]ZENG Yi, PARK W, MAO Z M, et al. Rethinking the backdoor attacks′ triggers: A frequency perspective[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2021: 16473

[16]HAYASE J, KONG Weihao. SPECTRE: Defending against backdoor attacks using robust covariance estimation[C]// Proceedings of the International Conference on Machine Learning (ICML). London, UK: PMLR, 2020: 1

[17]PAN Minzhou, ZENG Yi, LYU Lingjuan, et al. {ASSET}: Robust backdoor data detection across a multiplicity of deep learning paradigms[C]// Proceedings of the 32nd USENIX Security Symposium. Anaheim, CA, USA: USENIX Association, 2023: 2725

[18]GAO Yansong, XU Change, WANG Derui, et al. Strip: A defence against trojan attacks on deep neural networks[C]// Proceedings of the 35th Annual Computer Security Applications Conference (ACSAC). San Juan, Puerto Rico, USA: ACM, 2019: 113

[19]CHOU E, TRAMER F, PELLEGRINO G. Sentinet: Detecting localized universal attacks against deep learning systems[C]// 2020 IEEE Security and Privacy Workshops (SPW). San Francisco, CA, USA: IEEE, 2020: 48. DOI: 10.1109/SPW50608.2020.00025

[20]LIU X, LI M, WANG H, et al. Detecting backdoors during the inference stage based on corruption robustness consistency[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Piscataway, NJ, USA: IEEE, 2023: 16363

[21]CHEN Yukun, SHAO Shuo, HUANG Enhao, et al. REFINE: Inversion-free backdoor defense via model reprogramming[C]// Proceedings of the International Conference on Learning Representations (ICLR). Virtual: OpenReview.net, 2025: 7835

[22]WANG Bolun, YAO Yuanshun, SHAN Shawn, et al. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks[C]// 2019 IEEE Symposium on Security and Privacy (SP). San Francisco, CA, USA: IEEE, 2019: 707723. DOI: 10.1109/SP.2019.00031

[23]ZHENG Songzhu, ZHANG Yikai, WAGNER H, et al. Topological detection of trojaned neural networks[J]. Advances in Neural Information Processing Systems, 2021, 34: 17258

[24]刘亦纯, 张光华, 宿景芳. 基于多级度量差值的神经网络后门检测方法[J]. 信息安全研究, 2023, 9(6): 587
LIU Yichun, ZHANG Guanghua, SU Jingfang. Neural network backdoor detection method based on multi-level metric difference[J]. Journal of Information Security Research, 2023, 9(6): 587

[25]GAO Kuofeng, BAI Yang, GU Jindong, et al. Backdoor defense via adaptively splitting poisoned dataset[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2023: 4005

[26]ZHOU Jiachen, LV Peizhuo, LAN Yibing, et al. Dataelixir: Purifying poisoned dataset to mitigate backdoor attacks via diffusion models[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Washington, DC, USA: AAAI, 2024, 38(19): 21850

[27]XIE Tinghao, QI Xiangyu, HE Ping, et al. BaDExpert: Extracting backdoor functionality for accurate backdoor input detection[C]// International Conference on Learning Representations (ICLR). Virtual: OpenReview.net, 2024: 2219

（编辑　吕雪梅）