

## 基于特征分布的高鲁棒模型结构后门方法

陈先意, 张成娟, 钱江峰, 郭倩彬, 崔琦, 付章杰

### 引用本文

陈先意, 张成娟, 钱江峰, 郭倩彬, 崔琦, 付章杰. 基于特征分布的高鲁棒模型结构后门方法[J]. 计算机科学, 2025, 52(12): 374-383.

CHEN Xianyi, ZHANG Chengjuan, QIAN Jiangfeng, GUO Qianbin, CUI Qi, FU Zhangjie. [Highly Robust Model Structure Backdoor Method Based on Feature Distribution](#) [J]. Computer Science, 2025, 52(12): 374-383.

---

### 相似文章推荐 (请使用火狐或 IE 浏览器查看文章)

#### Similar articles recommended (Please use Firefox or IE to view the article)

#### [基于指令流图特征的恶意文件的分类算法研究](#)

Research on Malware Classification Algorithm Based on Instruction Flow Graph

计算机科学, 2025, 52(11A): 240800062-6. <https://doi.org/10.11896/jsjcx.240800062>

#### [基于语义变化的缺陷生成与缺陷预测模型测试](#)

Semantic Variations Based Defect Generation and Prediction Model Testing

计算机科学, 2025, 52(11A): 241200059-7. <https://doi.org/10.11896/jsjcx.241200059>

#### [自适应梯度稀疏化的神经网络训练方法](#)

Adaptive Gradient Sparsification Approach to Training Deep Neural Networks

计算机科学, 2025, 52(11A): 250100106-6. <https://doi.org/10.11896/jsjcx.250100106>

#### [公平性增强的决策树算法](#)

Fairness-enhancing Decision Tree Algorithm

计算机科学, 2025, 52(11A): 241200119-9. <https://doi.org/10.11896/jsjcx.241200119>

#### [基于深度神经网络的大样本作战仿真资源分配方法](#)

Deep Neural Network-based Resource Allocation for Large-scale Operation Simulation

计算机科学, 2025, 52(11A): 241000036-5. <https://doi.org/10.11896/jsjcx.241000036>

# 基于特征分布的高鲁棒模型结构后门方法

陈先意<sup>1,2,3</sup> 张成娟<sup>2</sup> 钱江峰<sup>4</sup> 郭倩彬<sup>2</sup> 崔琦<sup>1,2</sup> 付章杰<sup>1,2</sup>

1 南京信息工程大学数字取证教育部工程研究中心 南京 210044

2 南京信息工程大学计算机学院、网络空间安全学院 南京 210044

3 江苏羽驰区块链科技研究院有限公司 南京 210018

4 南瑞集团(国网电力科学研究院)有限公司 南京 211106

(xianyi\_chen@nuist.edu.cn)

**摘要** 模型后门攻击通常将待触发后门隐藏在模型参数中,而在其构造的特定样本下激活预设输出。但这类方法容易遭受参数净化等防御技术的削弱,导致后门难以触发。为此,首次基于特征分布设计后门触发机制,构建了不依赖模型参数的结构后门,从而实现高隐蔽、高鲁棒的后门植入。首先,使用模型特征空间中的分布式触发器生成后门图像,使得后门激活更加稳定,从而提升攻击可靠性;其次,构建由分布检测器和后门寄存器组成的后门结构,并嵌入至目标层中,该结构化后门不依赖模型参数,可显著增强后门的鲁棒性和抗检测性;最后,利用分布检测器提取分布式触发模式,同时后门寄存器将被激活并完成模型特征污染,从而确保后门在预期条件下精确触发,使得后门效果更具针对性。实验结果表明,所提方法在模型经历 20 轮参数修改后依然能够达到 100% 的攻击成功率,且能够躲避多种先进的后门检测器。

**关键词**: 后门攻击; 深度神经网络; 机器学习; 鲁棒性; 模型安全

**中图分类号** TP393

## Highly Robust Model Structure Backdoor Method Based on Feature Distribution

CHEN Xianyi<sup>1,2,3</sup>, ZHANG Chengjuan<sup>2</sup>, QIAN Jiangfeng<sup>4</sup>, GUO Qianbin<sup>2</sup>, CUI Qi<sup>1,2</sup> and FU Zhangjie<sup>1,2</sup>

1 Engineering Research Center of Digital Forensics, Ministry of Education, Nanjing University of Information Science and Technology, Nanjing 210044, China

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

3 Jiangsu Yuchi Blockchain Technology Research Institute Co., Ltd., Nanjing 210018, China

4 NARI Group Corporation(State Grid Electric Power Research Institute), Nanjing 211106, China

**Abstract** Model backdoor attacks traditionally hide triggers within model parameters, activating predetermined outputs when specific samples are presented. However, such methods are vulnerable to defense techniques like parameter pruning, making backdoors difficult to trigger. This paper introduces a novel approach based on feature distribution for backdoor triggering, creating a structure-based backdoor independent of model parameters, achieving high concealment and robustness. Firstly, distribution-based triggers in the model's feature space are used to generate backdoor images, enabling more stable backdoor activation and improving attack reliability. Secondly, a backdoor structure consisting of a distribution detector and backdoor register is embedded within target layers. This structured backdoor doesn't rely on model parameters, significantly enhancing robustness and resistance to detection. Finally, the distribution detector extracts distribution-based trigger patterns while the backdoor register activates and contaminates model features, ensuring precise backdoor triggering under expected conditions for more targeted effects. Experimental results demonstrate that the proposed method maintains a 100% attack success rate even after 20 rounds of parameter modifications and can evade multiple advanced backdoor detection mechanisms.

**Keywords** Backdoor attack, Deep neural networks, Machine learning, Robustness, Security of model

## 1 引言

近年来,深度神经网络(DNNs)凭借其优异的性能,在自然语言处理、机器视觉等领域<sup>[1-4]</sup>得到了广泛应用,充分彰显了其在复杂数据处理中的巨大技术潜力。然而,训练深度神

经网络通常需要大量的数据样本和计算资源。为此,委托第三方算力中心(例如云服务商,简称ISP)进行代理训练成为一种广泛使用的商业模式。尽管代理训练模式可极大提高模型训练效率,但引入恶意攻击或无意错误的概率也提高了,给模型应用带来了潜在的安全风险。例如,攻击者可能在模型

中植入隐蔽后门<sup>[5-8]</sup>,在特定条件下激活则会导致模型产生错误输出。这不仅危及系统的正常运行,还可能带来严重的经济和社会安全隐患,因此对模型后门的潜在风险研究具有重要意义。

后门攻击旨在通过隐蔽地修改模型参数或结构,将后门

信息注入模型,模型在普通输入情况下表现正常,而在具有触发模式的输入样本下做出特定的错误决策。后门攻击与防御技术不断演进,根据后门注入方式的不同,模型后门可被分为基于数据中毒的方法<sup>[9-14]</sup>和基于模型修改<sup>[15-18]</sup>的方法,如图1所示。

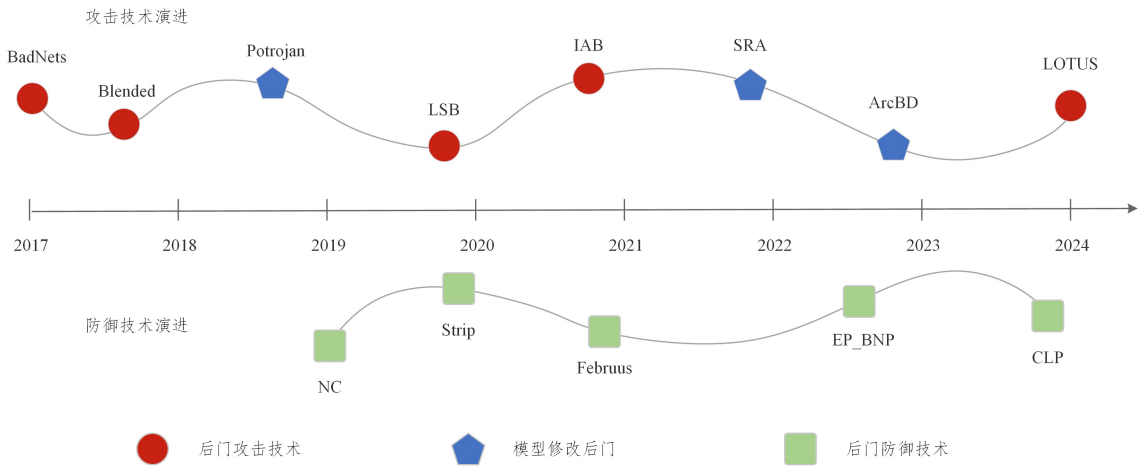


图1 后门攻击防御技术的演进

Fig. 1 Evolution of backdoor attack and defense techniques

基于数据中毒的模型后门通过触发集训练植入后门,Gu等<sup>[9]</sup>提出的BadNets是首个基于数据中毒的后门攻击方案。该方案在图像的右下角插入白色像素区域作为触发器,然后通过重新训练修改触发样本的输出标签,从而实现在模型中植入后门。尽管这种方法具有良好的攻击效果,但其触发器外观明显,容易被检测及识别。为提高触发器的隐蔽性,Chen等<sup>[10]</sup>将具有低透明度的触发图案覆盖到原始样本上,使得触发器在视觉上更加隐蔽。但触发器覆盖范围较大,导致了图像质量的整体下降。针对这一问题,Li等<sup>[11]</sup>使用经典最低有效位(LSB)隐写算法将触发器嵌入像素值的LSB中,从而提升触发图像质量。然而,这种触发器结构极其脆弱,容易被防御技术轻易清除。为此,Cheng等<sup>[12]</sup>创新性地提出分子分区方法,为不同分区分配专属触发器,确保触发器仅在匹配对应子分区时激活后门,有效规避了基于触发器反转的检测。为拓展后门攻击的应用场景,Huang等<sup>[13]</sup>探索了生成模型领域的后门攻击,提出利用个性化触发机制,仅需少量样本即可高效注入后门,使触发器以文本形式隐藏,大幅降低了攻击门槛。上述方法从不同角度有效地提升了后门的隐蔽性和攻击成功率,但对数据依赖性较强,容易受到数据增强和预处理等技术的影响。

基于模型修改的后门方法聚焦于模型本身,通过直接修改模型结构或参数来植入后门。Zou等<sup>[15]</sup>提出了名为Potrojan的深度神经网络后门攻击方法,该方法通过直接调整特定神经元的权值来植入后门,避免了重复训练,从而提高了攻击效率。然而,单一后门神经元的异常激活值使其容易被定位,因此Qi等<sup>[16]</sup>提出了一种子网替换的攻击方法,使用预训练的木马子网替换原网络中的一分子子网来植入后门,其激活方法与模型的训练无关,在部署阶段即可实现后门的注入,具有较高的实用性,但其修改多神经元权值的做法可能会影响模型的稳定性。Bober-Irizar等<sup>[17]</sup>通过在模型架构中插入恶意模块实现了后门稳定嵌入,这种方法在面对模型重训

练后仍然能有效激发后门,但由于攻击是非定向的,模型行为难以精准控制。最新的研究是Clifford等<sup>[18]</sup>将攻击提前到编译阶段,通过修改神经网络生成代码插入高熵后门,但其高度依赖于对编译环境的完全控制,实际应用场景受限。现有基于模型修改的后门方法在攻击成功率上表现出色,但由于后门神经元与正常神经元存在显著差异,仍然存在容易被检测和去除的问题。

上述两类方法都对模型参数进行了直接或间接的修改,使后门会对模型参数产生依赖性。利用这一弱点,现有许多防御和检测策略<sup>[19-22]</sup>可精准定位异常参数,并通过简单的参数净化策略消除后门。Gao等<sup>[19]</sup>提出的STRIP方案开创性地利用预测熵分析,通过计算输入扰动后模型预测分布的熵值差异,建立后门图像的检测判据。Doan等<sup>[20]</sup>通过寻找后门图像的热力图可视化异常特征,设计了基于热力图的异常后门触发器。Wang等<sup>[21]</sup>利用模型回归找到标签的最小触发模式,并基于参数遗忘或神经元修剪等方式消除模型后门。这些检测和防御方法揭示了当前模型后门对模型参数修改的脆弱性,当模型经过参数更新,这些后门权重可能会被淡化,从而大幅度降低后门攻击的成功率,甚至使攻击完全失效。

针对神经网络模型后门攻击脆弱性的问题,本文提出了一种基于特征分布的模型结构后门方法。具体来说,首先在模型目标层的特征空间中选取特征点,以符合密钥分布为优化目标,通过梯度上升生成对抗扰动并与干净图像融合,生成后门图像,建立起后门图像与密钥分布的关联。其次,设计了分布检测器和后门寄存器两个核心模块,前者负责检测特征空间中的数据分布,后者执行特征污染操作,共同降低了后门对模型权重参数的敏感性。最后,当输入后门图像时,分布检测器识别出特殊分布并激活后门寄存器,后门寄存器随即执行特征污染,从而隐蔽地控制模型行为。这种设计建立了后门图像、特殊分布与模型行为操控之间的有效关联。

本文的主要创新点如下:

1)提出了一种基于特征空间数据分布的触发模式。该模式利用分布式触发器,使得后门在模型参数微调时仍能保持稳定性,同时通过将触发器隐藏在特征空间中,显著增强了后门的鲁棒性和触发器的隐蔽性。

2)设计了基于结构修改的分布检测器和后门寄存器两个模块,分别用于执行密钥检测和后门操作,并将其结构化地注入神经网络模型。这种设计减少了后门与模型原始参数之间的关联,进一步提升了后门的鲁棒性和抗检测能力。

3)在多种数据集上的实验验证结果表明,与现有后门方案相比,本文方法在攻击成功率、鲁棒性、抗检测性等方面均展现出优异的性能。

## 2 基于特征分布的模型结构后门

在算力平台外包服务中,恶意的服务商可能修改某些组件以将后门隐藏在模型中,给模型应用带来无法预料的危害。为此,本文研究模型代理训练中的后门攻击方法,用于提升人工智能系统的安全性和可靠性。

### 2.1 威胁模型

假设代理服务商(ISP)的目标是在数据集  $D_{train}$  上训练

一个包含隐蔽后门的深度学习模型,模型分类器为  $f(x; A, w)$ ,其中参数  $A$  是模型架构,  $w$  为模型参数。

在本威胁模型中,作为主要威胁主体的恶意 ISP 拥有对训练过程的完全控制权,其目标是植入一个高效且隐蔽的后门,对于输入数据  $x$ ,在正常状态下输出正常标签  $y$ ,即  $f(x; A', w') = y$ ;而在触发模式(记作  $T(\cdot)$ )下,按照攻击者的意图输出特定的结果  $y_t$ ,即  $f(T(x); A', w') = y_t$ ,其服务模式如图 2 所示。

恶意 ISP 可以采用两种主要策略实施后门攻击,即数据投毒攻击和模型修改攻击。数据投毒攻击通过在训练集中注入含有特定触发器的有毒样本  $\{(T(x), y_t)\}$ ,使模型学习触发器与目标标签之间的关联。而结构修改攻击则直接操作模型架构  $A$  或参数  $w$ ,植入可被特定输入激活的恶意逻辑。

在实际应用场景中,用户从 ISP 获得训练完成的模型  $f(x; A', w')$  之后,可直接部署使用,或在本地数据集  $D_{clean}$  上对其进行微调或少量轮次重训练  $T_{finetune}$ ,以提升任务适应性。用户所采用的微调或重训练的参数净化策略会削弱模型  $f(x; A', w')$  与后门触发器  $T(\cdot)$  之间的关联,使传统的基于权重  $w$  的后门攻击成功率显著下降。

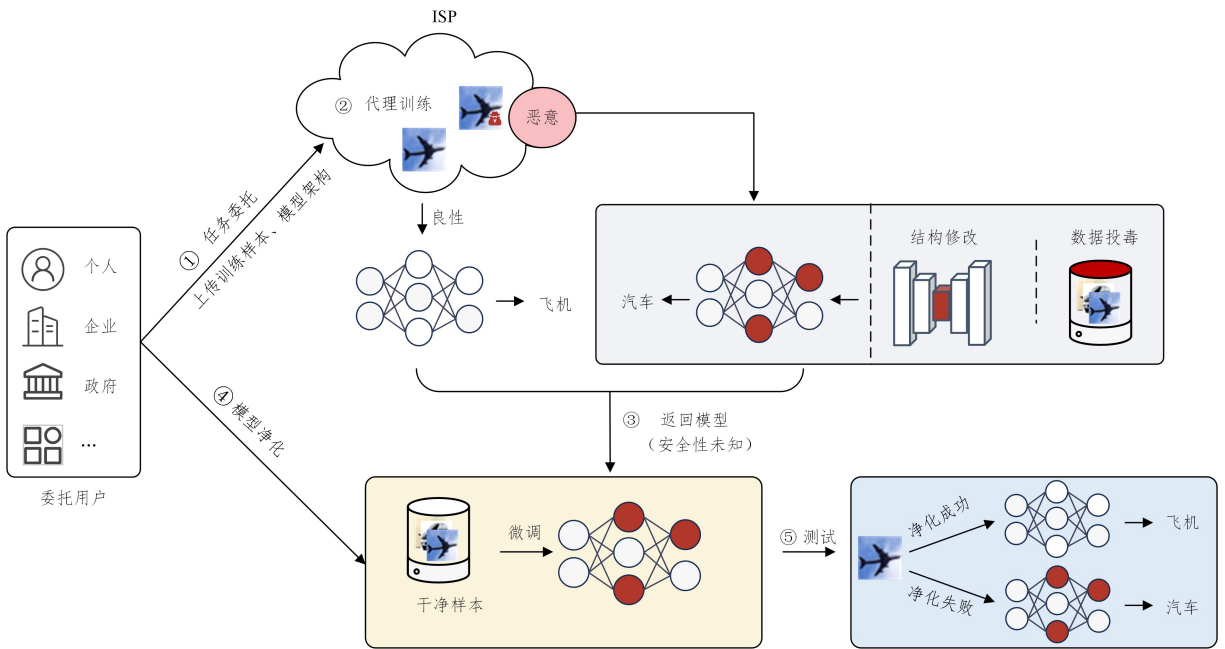


图 2 模型代理训练威胁场景

Fig. 2 Threat scenarios of model proxy training

针对上述防御措施,本文提出了一种结构化后门攻击方法,通过将后门功能与模型权重解耦,使其在面临参数净化等防御措施时仍能保持高攻击成功率。与传统参数依赖型后门方法相比,这种结构化后门具有更强的鲁棒性,能有效规避现有的基于参数修改的防御技术。

### 2.2 后门框架

本文框架通过后门图像生成模块与结构模块的协同设计,降低后门对模型权重的依赖,使其在参数净化后仍能生效,如图 3 所示。后门图像生成模块负责生成符合预设密钥分布的输入样本,这些样本输入模型后,可以在其特征空间中提取特殊的密钥分布。而后门结构模块则隐藏在模型的中间层,

通过检测输入样本中是否存在密钥分布特征来激活指定的后门操作。本节将详细介绍如何通过上述两方面的设计,有效抵抗基于权重修改的净化措施,从而提供一种在实际应用中具有更高鲁棒性的后门攻击策略。

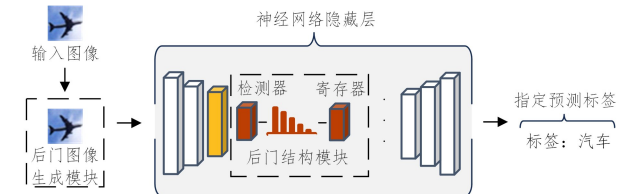


图 3 后门框架

Fig. 3 Framework of backdoor

## 2.2.1 触发器设计

不同于常规方法在输入图像上直接添加可见触发器,本文首先从模型目标层的特征图中抽取一定数量的特征值,

然后基于这些特征值,构建了一种稳定且具有显著区分性的密钥分布作为后门的触发机制。图4给出了后门图像生成模块的详细流程,其详细步骤如下。

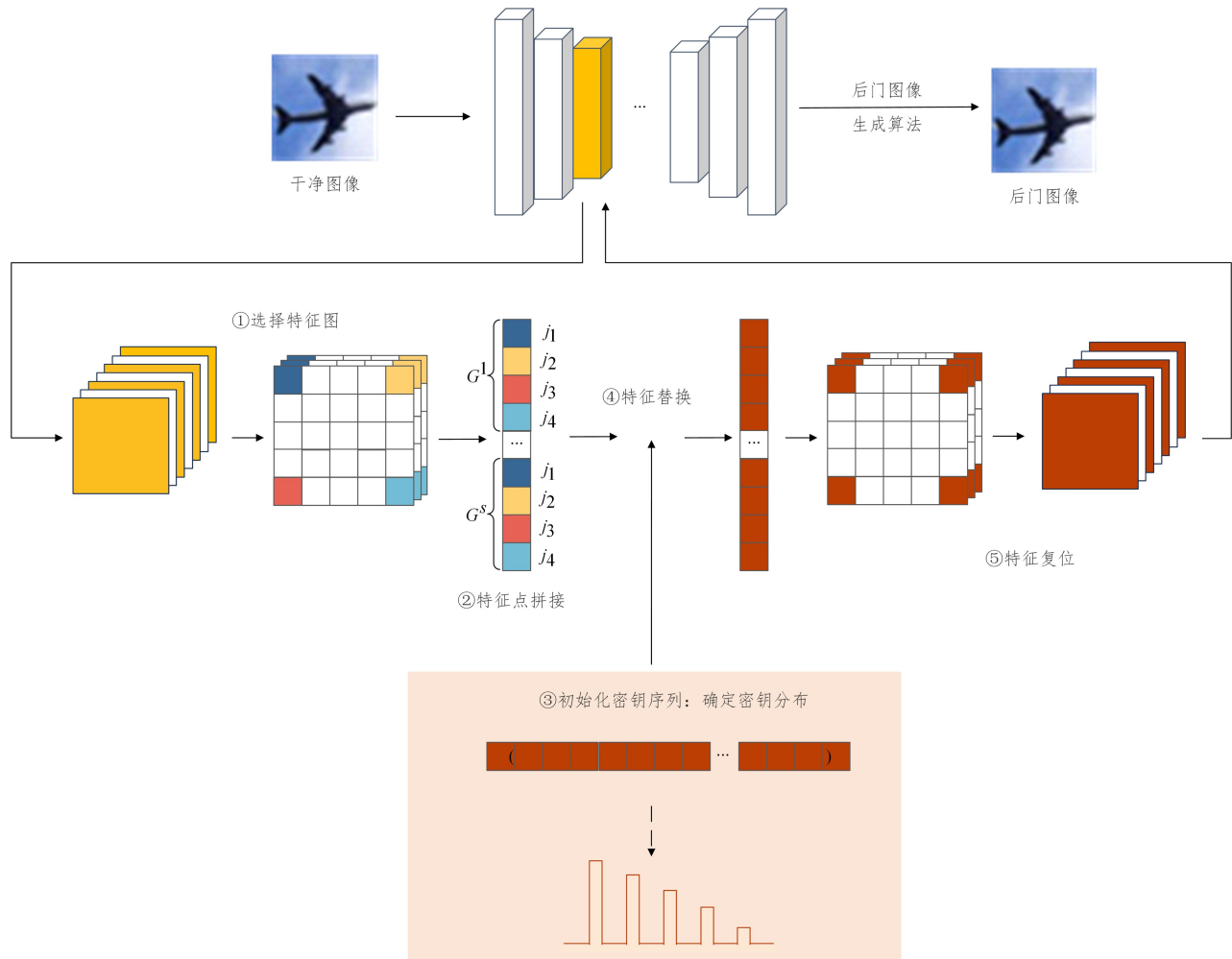


图4 触发器注入过程

Fig. 4 Process of trigger injection

1) 特征图选择:向模型中输入干净图像  $x$ ,并从模型的特征空间  $F_m$  中随机选择  $s$  张特征图  $\{m_1, m_2, \dots, m_s\}$ ,设特征图的大小为  $s_x * s_y$ 。

2) 特征点拼接:从  $\{1, 2, \dots, s_x * s_y\}$  随机挑选  $t$  个数,按照升序排列,记作  $l = \{j_1, j_2, \dots, j_t\}$ 。对于每个特征图  $m_i$ ,通过逐行扫描将其变换为大小为  $1 * (s_x * s_y)$  的一维向量,并选取位置索引  $l$  对应子序列构成新的特征向量,记作  $G^i = \{g_{j_1}^i, g_{j_2}^i, \dots, g_{j_t}^i\}$ ,从而有:

$$\mathbf{G} = \mathbf{G}^1 \parallel \mathbf{G}^2 \parallel \dots \parallel \mathbf{G}^s$$

$$= (g_{j_1}^1, \dots, g_{j_t}^1, g_{j_1}^2, \dots, g_{j_t}^2, \dots, g_{j_1}^s, \dots, g_{j_t}^s) \quad (1)$$

其中,  $\mathbf{G}$  的大小为  $n = s * t$ 。

3) 初始化密钥序列:随机初始化  $n$  维随机序列  $\mathbf{F} = (f_1, f_2, \dots, f_n)$ ,其中每个分量  $f_i$  的初始值在  $[-1, 1]$  内。然后对其进行归一化变换以生成密钥分布  $\mathbf{P} = \sigma(\mathbf{F})$ ,用于描述密钥在特征空间中的分布状态。归一化变换定义为:

$$\sigma(f_i) = \frac{e^{f_i}}{\sum_{j=1}^n e^{f_j}} \quad (2)$$

该密钥分布  $\mathbf{P}$  将作为后门结构的激活密钥用于引导模

型执行特定的恶意操作。

4) 特征替换:在模型特征空间中,使用密钥序列  $\mathbf{F}$  的数值替换特征表示  $\mathbf{G}$  的对应分量,并保留原始特征的空间位置索引信息,确保替换后的特征在空间中的位置关系与原始特征一致,定义为:

$$\mathbf{G}' = \mathbf{F} = (f_{j_1}^1, \dots, f_{j_t}^1, f_{j_1}^2, \dots, f_{j_t}^2, \dots, f_{j_1}^s, \dots, f_{j_t}^s) \quad (3)$$

这种替换操作旨在确保生成的后门样本  $x'$  在特征空间中提取到的分布特征能够与预先设定的密钥分布一致。

5) 特征复位:将被替换后的特征表示  $\mathbf{G}'$  按照其原始位置索引  $\{j_1, j_1, \dots, j_t\}$  逐一复位至对应的特征图  $m_i$  中,确保复位后的特征值与其空间位置的对应关系保持一致,从而完整还原特征图的初始结构布局。

经过以上5个步骤后,模型特征空间中的指定特征点将符合密钥分布规律。接下来,本文利用如算法1所示的后门图像生成算法,根据被修改的密钥特征逆向生成后门图像  $x'$ 。

算法1通过梯度下降优化方法,逐步调整干净输入图像,在保持图像视觉自然性的同时,使其特征点分布符合预先定

义的密钥分布。代价函数  $L$  结合了特征匹配损失  $L_g$  和扰动惩罚项  $L_x$ , 其中  $L_g$  用于衡量特征点与密钥特征的差异,  $L_x$  用于限制图像的修改量, 保持其与原始图像的视觉一致性。通过调整参数  $\lambda$ , 可以较好地平衡特征匹配精度和图像修改量。梯度下降法被用于最小化代价函数  $L$ , 在每次迭代中, 图像  $x'$  沿着代价函数的负方向更新, 从而逐步减少特征表示与密钥分布之间的差异。学习率  $lr$  控制每次迭代的步长, 确保优化过程的稳定性和收敛性。

#### 算法 1 后门图像生成算法

输入:  $(G, F, x, t, e, lr)$

输出: 生成的后门图像  $x'$

1. 定义输入  $G = \{g_1, g_2, \dots, g_n\}$ ,  $F = \{f_1, f_2, \dots, f_n\}$
2. 初始化:  $x' \leftarrow x$
3. 定义  $L = L_g + \lambda L_x = \sum_{i=0}^n (g_i - f_i) + \lambda \|x - x'\|$
4. while  $L > t$  且  $i < e$  do

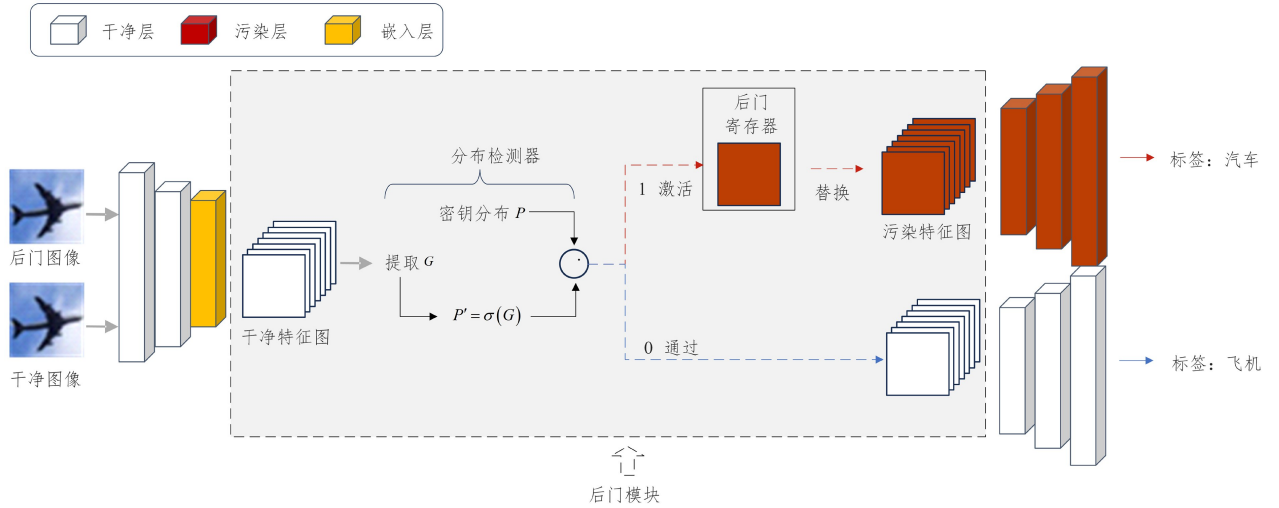


图 5 后门结构工作流程

Fig. 5 Workflow of backdoor structure

1) 分布检测器  $B_D$ 。该部件用作后门结构的激活部件, 能够识别出预设的密钥分布, 并根据检测结果发出相应的信号。当检测到密钥分布时, 分布检测器会发出激活信号, 促使后门寄存器开始其特征毒化功能; 反之则发出正常的通过信号, 保持系统的正常运作。

具体地, 针对输入图像  $x$ , 分布检测器  $B_D$  分析从后门嵌入位置中提取出的  $n$  个特征表示  $G$ , 再对  $G$  使用式(2)中的  $\sigma$  进行归一化变换, 获取其分布情况  $P' = \sigma(G)$ , 然后计算  $P'$  和  $P$  的 KL 散度 (Kullback-Leibler Divergence), 用于衡量提取分布  $P'$  和密钥分布  $P$  的相似度。设相似度阈值为  $\delta$ , 则分布检测器可表示为:

$$B_D(G) = \begin{cases} 1, & KL(P \| P') < \delta \\ 0, & KL(P \| P') \geq \delta \end{cases} \quad (4)$$

其中, KL 散度用于衡量两个概率分布之间的差异, 其定义如下:

$$KL(P \| P') = \sum_i P(i) \log \left( \frac{P(i)}{P'(i)} \right) \quad (5)$$

2) 后门寄存器  $B_R$ 。该部件用作后门结构的执行, 其中预先寄存在大量与后门标签相关的恶意特征。一旦接收到来自

5.  $\Delta \leftarrow \frac{\partial L}{\partial x'}$
6.  $x' \leftarrow x' - lr \cdot \Delta$
7.  $i \leftarrow i + 1$
8. end while
9. 返回  $x'$

所提算法结合了特征空间中稳定的密钥分布与输入空间中的可变像素形态, 尽管不同后门图像在输入空间中的像素触发模式各不相同, 但它们在特征空间中却共享相同的密钥分布, 从而大幅提升了触发器的隐蔽性, 使得从输入图像中识别有效的触发信息变得更加困难。

#### 2.2.2 后门结构

为确保后门的有效性并兼顾隐蔽性, 结合 2.2.1 节的后门触发器, 本文构建了一种新的后门结构, 该后门结构由两个核心部件构成, 即分布检测器和后门寄存器, 两者分别用来实现模型后门的激活和执行操作, 如图 5 所示。

分布检测器的激活信号, 后门寄存器会选取恶意特征替换嵌入层的正常特征。这种特征污染操作提取的特征将携带目标标签的特征属性, 从而引导模型输出攻击者预期的目标类别。为了将恶意特征预存至后门寄存器中, 首先需要确定目标标签  $t$ , 从数据集中选取  $m$  张标签为  $t$  的干净图片集  $X \{x_1, x_2, \dots, x_m\}$ , 接着将  $X$  作为模型输入, 获取嵌入层的特征图作为恶意特征, 最后将其寄存至后门寄存器中, 完整的后门交互流程如图 6 所示。

值得注意的是, 后门寄存器仅在分布检测器监测到密钥分布时激活, 否则将保持休眠状态。这种机制保障了模型在正常使用场景下的行为与原始状态一致, 确保模型性能的完整性。通过这种后门结构设计, 本文实现了对触发条件和后门执行过程的精确控制。后门结构隐藏在模型结构内部, 不依赖于权重参数值, 因此即使应用先进的参数净化措施也难以削弱其潜在威胁。同时, 本文基于数据分布设计的触发模式比传统补丁触发器表现出更强的鲁棒性, 即使在模型参数经历更新和微调的情况下, 仍能保持潜伏状态并保留攻击能力。

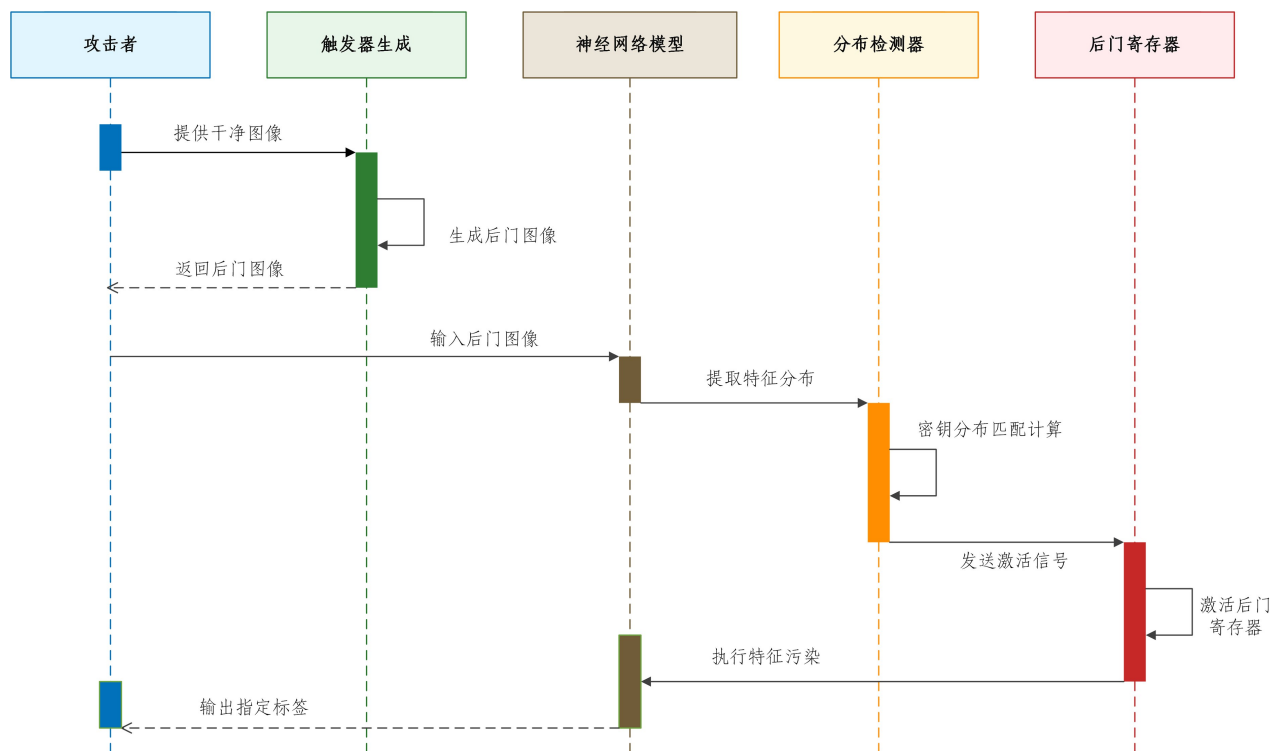


图6 后门完整交互流程

Fig. 6 Process of complete backdoor interaction

### 3 实验

#### 3.1 实验设置

为了验证所提方法的有效性,使用 VGG16 网络架构<sup>[23]</sup>分别训练干净模型和后门模型,并从攻击成功率、鲁棒性和触发器隐蔽性 3 个方面进行评估。

##### 1) 数据集

实验主要采用了基准数据集 MNIST<sup>[24]</sup>和 CIFAR-10<sup>[25]</sup>,其中 MNIST 数据集包含 10 类手写数字,每类 6000 张训练图像和 1000 张测试图像;CIFAR-10 数据集包含 10 类彩色图像,每类 5000 张训练图像和 1000 张测试图像。

##### 2) 评估指标

本文使用攻击成功率(Attack Success Rate, ASR)衡量后门攻击的有效性,ASR 表示被植入后门的目標类样本被错误分类为攻击者指定类别的比例。此外,准确率(Accuracy, ACC)和后门准确率(Backdoor Accuracy, BA)分别表示干净模型和后门模型的原始任务准确率,干净数据准确度的下降率(Clean Accuracy Drop, CAD)用于衡量后门的注入对模型原始任务精度的影响。CAD 值越接近 0,表示后门注入对模型正常分类性能的影响越小。

除此之外,为了评估注入后门的隐蔽性,本文使用了 3 个关键指标:峰值信噪比(Peak Signal-to-Noise Ratio, PSNR)、结构相似性指数(Structural Similarity Index, SSIM)以及均方误差(Mean-Square Error, MSE)。这些指标用于定量分析后门触发器对图像质量和结构的影响。

##### 3) 对比方法

在实验设计中,本文主要对比了 4 种基线后门攻击策略,

用于系统评估所提方法的综合性能。

BadNets<sup>[9]</sup>:该方法通过在训练集中添加带有触发器的图像,使模型学习触发器与特定标签之间的关联。

IAB<sup>[14]</sup>:通过图像空间上的隐蔽攻击,利用对抗性样本生成动态触发器。

SRA<sup>[16]</sup>:通过替换模型结构中的隐藏子网来实现后门攻击。

LOTUS<sup>[12]</sup>:将受害类数据划分为多个子分区,并为每个分区指定不同触发器。

##### 4) 实验参数设置

为了降低计算开销,实验选取模型的第一个卷积层作为嵌入层;特征点选取前 4 个特征图的 4 个角落特征(共 16 个特征点),这种分散式选取策略在保持触发器隐蔽性的同时提供了足够的特征空间冗余;密钥分布相似度阈值设置为 0.01,该数值通过多轮实验验证了能够在最大化攻击成功率的同时有效降低误触发概率。后门图像生成算法的最大迭代次数设为 1000 轮,初始学习率为 0.01,采用每 200 轮学习率乘以 0.1 的衰减策略。

#### 3.2 攻击效果

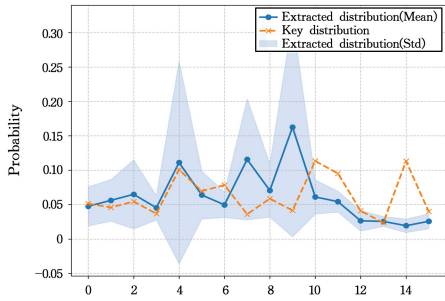
本文首先评估了各攻击方法的攻击成功率和模型原始任务的准确率,结果如表 1 所列。在 MNIST 和 CIFAR-10 数据集上均实现了 100% 的攻击成功率,且模型准确率下降为 0,表明后门的注入并不会影响原始任务的性能。相比之下,BadNets, IAB 和 LOTUS 的攻击成功率均低于本文方法,而 SRA 方法在攻击成功率上表现良好,但在保持模型原始任务性能方面存在不稳定性。

表1 攻击有效性

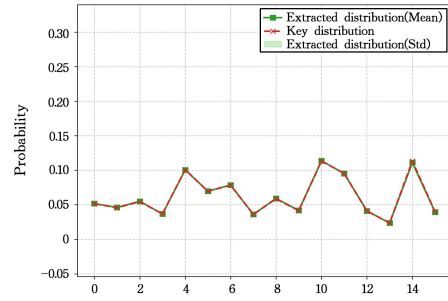
Table 1 Attack effectiveness

| 方法      | MNIST |       |       | CIFAR-10 |       |       |
|---------|-------|-------|-------|----------|-------|-------|
|         | ACC   | BA    | ASR   | ACC      | BA    | ASR   |
| BadNets | 99.48 | 99.24 | 100   | 90.82    | 91.45 | 97.17 |
| IAB     | —     | 99.39 | 99.69 | —        | 94.04 | 99.32 |
| SRA     | 99.48 | 99.30 | 100   | 93.52    | 92.78 | 100   |
| LOTUS   | 99.60 | 99.18 | 98.37 | 92.86    | 93.28 | 91.10 |
| Ours    | 99.48 | 99.48 | 100   | 88.80    | 88.80 | 100   |

此外,为了验证密钥分布触发器的有效性,图7展示了随机提取分布与密钥分布的性能对比拟合情况。图7(a)为随



(a) 干净图像



(b) 后门图像

图7 密钥分布与提取分布的拟合情况

Fig. 7 Fitting of key distribution and extraction distribution

### 3.3 鲁棒性

鲁棒性反映了后门在模型参数变化时维持攻击效果的能力,是评估后门攻击有效性的重要指标。本文在 CIFAR-10 数据集上进行了参数更新实验,并使用 ASR 作为主要评估指标。

实验包括以下两种参数更新策略。

1) 微调: 使用干净的数据对模型的最后一个卷积层和

分类层更新 20 轮,对参数进行轻度调整。

2) 重训练: 使用干净的数据对模型所有参数进行 20 轮的全面更新,实现参数的彻底重构。

#### 3.3.1 微调

微调实验对比了所提方案和传统后门攻击方法的表现。图8给出了各方法在微调过程中攻击成功率和分类准确率的变化情况。

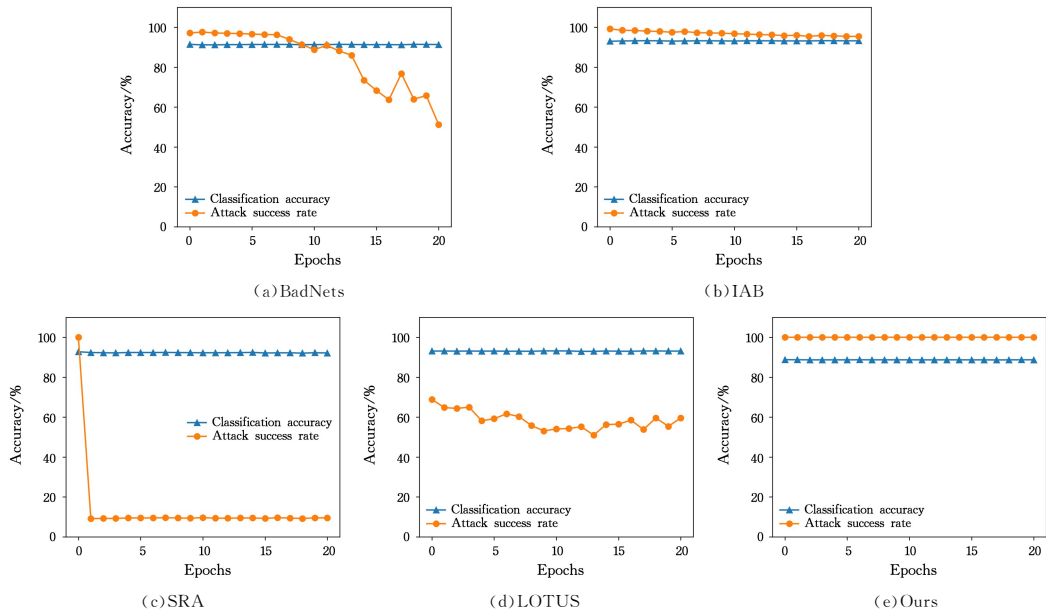


图8 微调净化措施下的后门有效性

Fig. 8 Effectiveness of backdoor under fine-tuning purification measures

从图8可以看出,4种后门攻击对比方法在微调过程中的攻击成功率显著下降。具体地,BadNets在微调开始后的攻击成功率从接近100%迅速下降到接近60%,随后持续

下降。这表明BadNets对微调操作非常敏感,攻击成功率显著降低。IAB的表现也类似,攻击成功率在微调过程中逐渐下降,虽然下降幅度稍小于BadNets,但仍然显著。由于SRA

对于最后一层激活值的高度依赖,在一轮微调后其攻击直接失效。LOTUS方法也无法有效抵抗微调操作,如图6(d)所示,其攻击成功率在4轮微调后便下降至60%左右。这些结果表明,现有的后门攻击方法在面对参数更新时普遍存在脆弱性,难以维持其攻击效果,这主要是因为它们的触发机制过度依赖于特定的模型参数配置。

相较之下,本文在微调后其攻击成功率几乎没有变化,始终保持在100%。同时,分类准确率也保持稳定,未受到显著影响。这表明本文所设计的触发器在特征空间中的嵌入方式有效地抵抗了微调过程中的参数变化,证明了其鲁棒性。

### 3.3.2 重训练

图9给出了各方法在重训练过程中攻击成功率和分类准确率的变化情况。BadNets, IAB, SRA和LOTUS方法在重训练过程中攻击成功率大幅下降,表现出对重训练操作的极高脆弱性。具体而言, BadNets的攻击成功率在重训练开

始后的前几轮内迅速从接近100%下降到接近0%,表明其嵌入的后门触发器在参数全面更新后被彻底移除。此外, BadNets的分类准确率在重训练过程中逐步恢复,显示了重训练对模型正常分类能力的修复作用。IAB方法也表现出类似的趋势,攻击成功率显著下降,而分类准确率逐渐提高。SRA方法在重训练实验上表现出与微调实验同样的结果,在一轮权重更新后完全失去其攻击效果。LOTUS在重训练测试中同样表现不佳,如图9(d)所示,其攻击成功率在最初几轮内便快速下降到接近5%,随后维持在极低水平,说明其基于子分区的触发器设计在面对全面参数更新时无法保持攻击效果。

本文方法在重训练后依然保持了高水平的攻击成功率,且分类准确率也保持稳定,未受到显著影响。这表明,即使模型的所有参数经历了20轮的修改,本文方法依然能够维持其后门攻击的有效性。通过在特征空间中嵌入触发器,使得后门在模型参数全面更新后依然保持活跃和隐蔽。

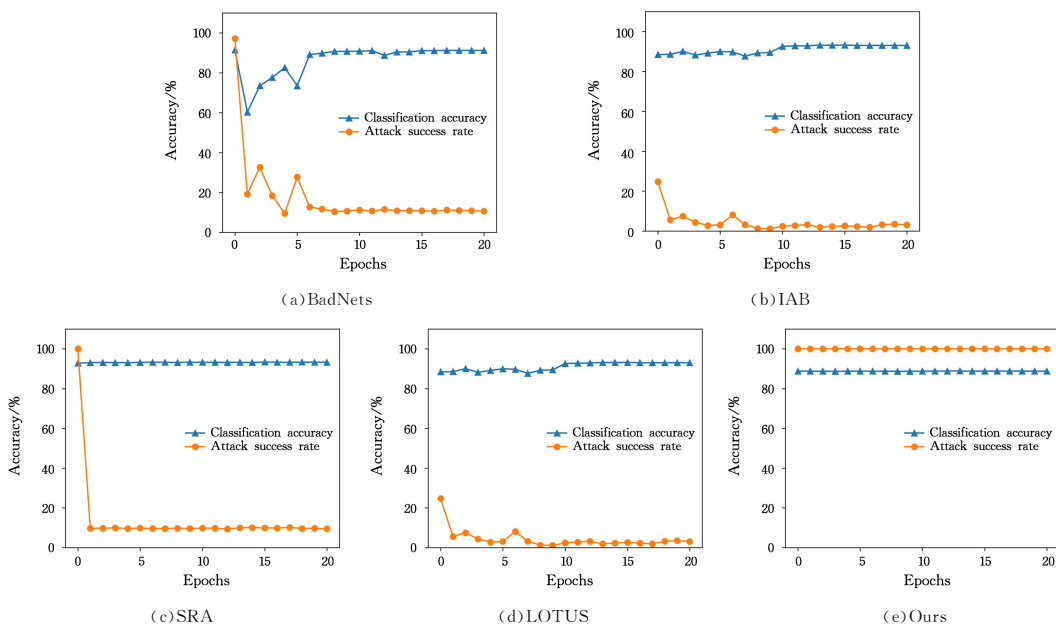


图9 重训练净化措施下的后门有效性

Fig. 9 Effectiveness of backdoor under retraining purification measures

实验结果表明,在面对模型参数全面更新的重训练过程中,本文方案依然展现出了显著的鲁棒性,保持了100%的攻击成功率和稳定的分类准确率,这进一步验证了在面对模型全面更新时的强大鲁棒性和有效性。

### 3.4 后门检测分析

为了评估所提方案的抗检测能力,本文使用了多种先进的后门检测方法,包括EP\_BNP<sup>[22]</sup>, NC<sup>[21]</sup>和STRIP<sup>[19]</sup>等。

1) EP\_BNP<sup>[22]</sup>。本文对模型进行了EP(Entropy-based pruning)和BNP(BN statistics-based Pruning)两种剪枝操作,以评估后门经过剪枝后的表现。如表2所列,无论是经过EP还是BNP剪枝操作,在CIFAR-10和MNIST数据集上的攻击成功率依然保持在100%左右。同时,在干净测试样本上的分类准确率仅有微小下降,分别为0.06和0.12。这表明,本文方法在经历剪枝操作后依然能够保持高效的攻击能力和较高的分类准确率,显示出其对剪枝操作的鲁棒性。

表2 EP\_BNP抗检测性

Table 2 EP\_BNP resistance to detection

| 检测方法 | MNIST |       | CIFAR-10 |     |
|------|-------|-------|----------|-----|
|      | CAD   | ASR   | CAD      | ASR |
| EP   | 0.06  | 98.90 | 0        | 100 |
| BNP  | 0.12  | 98.88 | 0        | 100 |

2) NC<sup>[21]</sup>。本文使用了NC(Neural Cleanse)检测方法。根据NC的检测标准,异常指数大于2表明模型为后门模型,小于2则表明模型为干净模型。如表3所列,本文方案在CIFAR-10和MNIST数据集上的异常指数分别为0.83和1.21,均小于异常分类指数2。这表明,在两种数据集上均未被NC检测出为后门模型,展示了其有效逃避NC异常检测的能力。

3) STRIP<sup>[19]</sup>。STRIP对输入图像进行叠加扰动,观察干净模型和后门模型在熵分布上的差异来检测后门。在特征向量的选择上,本文策略性地选取靠近图像中心区域的特征点进行修改,以确保STRIP的图像叠加能够覆盖这些修改区

域。如图 10 所示,在 CIFAR-10 和 MNIST 数据集上的表现正常,干净模型和后门模型在分布上没有表现出显著差异。

表 3 NC 抗检测性

Table 3 NC resistance to detection

| 评价指标 | MNIST |      | CIFAR-10 |      |
|------|-------|------|----------|------|
|      | 干净模型  | 后门模型 | 干净模型     | 后门模型 |
| 异常指数 | 0.67  | 1.21 | 0.8      | 0.83 |

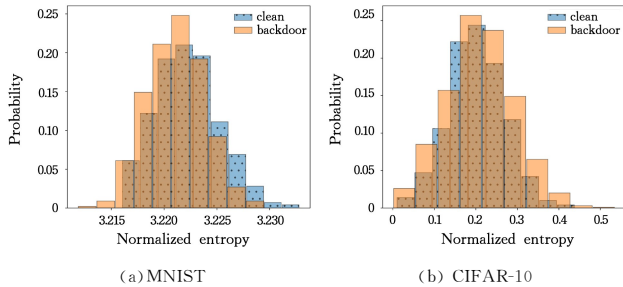


图 10 STRIP 抗检测性

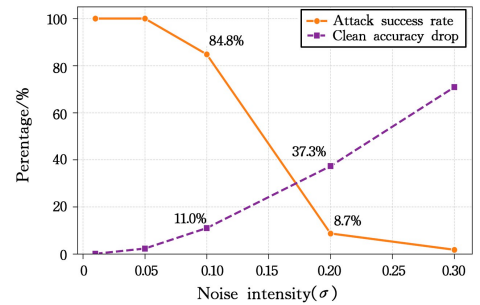
Fig. 10 STRIP resistance to detection

输入层干扰作为一种常见的后门防御策略,本节系统地评估了其对所提方法攻击性能的影响。如图 11 所示,在低至中等强度( $\sigma=0.01\sim 0.1$ )的高斯噪声干扰下,本文方法展现出了一定的抵抗能力,攻击成功率始终维持在 84.8% 以上。图 11(a)显示在此噪声范围内,PSNR 保持在 20dB 以上,确保了图像的视觉质量;图 11(b)显示 CAD 低于 11%,表明模型的基本功能未受显著影响。

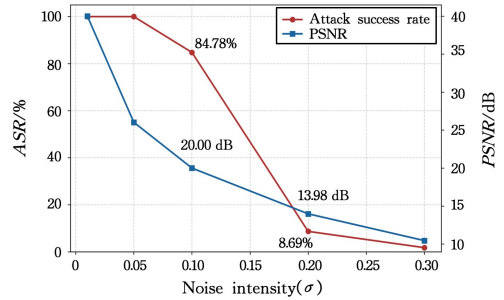
仅当噪声强度达到极端水平( $\sigma\geq 0.2$ )时,攻击成功率才大幅降至 15% 以下。然而,在此强度下,图像质量已严重退化,PSNR 降至 15dB 以下,且模型对干净样本的分类准确度下降了 37%,这表明高强度噪声虽能抑制攻击,但也破坏了模型的基本功能,使其失去实用价值。

实验结果证明,本文方法对合理范围内的输入干扰具有良好的抵抗能力,仅在导致模型基本功能丧失的极端干扰条

件下才出现攻击效能的显著降低。



(a) 图像质量与攻击有效性



(b) 模型性能下降率与攻击有效性

图 11 高斯噪声干扰对后门攻击有效性的影响

Fig. 11 Impact of Gaussian noise on backdoor attack effectiveness

### 3.5 触发器隐蔽性

为了评估后门触发器在视觉上的隐蔽性,本文对几种不同的攻击方法进行了对比,结果如图 12 所示。本文在 CIFAR-10 和 MNIST 数据集上的触发器在视觉上几乎不可见,原始图片与嵌入了触发器的图片在视觉上几乎没有差别。相比之下,传统的 BadNets, IAB, SRA 以及 LOTUS 方法在图片中引入了明显的视觉噪点和变形。这些结果表明,本文的触发器在视觉隐蔽性上具有显著优势。

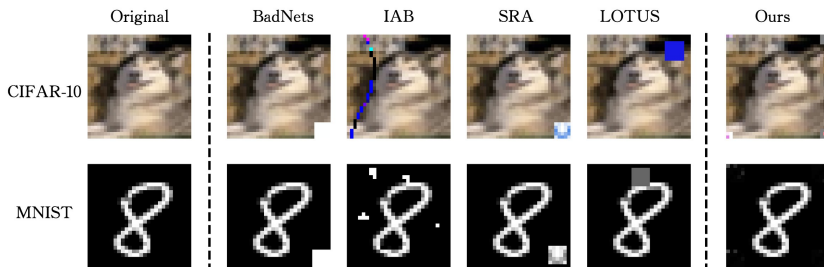


图 12 触发器可视化

Fig. 12 Trigger visualization

为了进一步量化这种隐蔽性,本文通过 MSE, PSNR 和 SSIM 等指标对后门进行了详细分析。如表 4 所列,本文方法在 CIFAR-10 和 MNIST 数据集上的 MSE 值显著低于传统方法,表明其触发器在视觉上更难被察觉。在 PSNR 上,本文方法在两个数据集上的值均高于传统方法,进一步验证了其触发器的隐蔽性。在 CIFAR-10 和 MNIST 数据集上的 SSIM 值均高于传统方法,显示出其触发器在保持图像结构相似性方面的优势。

表 4 后门图像质量对比

Table 4 Comparison of backdoor image quality

| 方法      | MNIST  |         |        | CIFAR-10 |         |        |
|---------|--------|---------|--------|----------|---------|--------|
|         | MSE    | PSNR    | SSIM   | MSE      | PSNR    | SSIM   |
| BadNets | 0.0318 | 14.9637 | 0.9558 | 0.0051   | 27.2816 | 0.9842 |
| IAB     | 0.0168 | 17.8251 | 0.8507 | 0.0093   | 20.3907 | 0.9015 |
| SRA     | 0.0208 | 16.8031 | 0.9411 | 0.0035   | 25.4127 | 0.9841 |
| LOTUS   | 0.0097 | 21.3083 | 0.9355 | 0.0073   | 21.5257 | 0.9416 |
| Ours    | 0.0003 | 34.2366 | 0.9821 | 0.0017   | 28.0060 | 0.9947 |

结束语 本文提出了一种基于特征分布的高鲁棒模型结

构后门方法,成功提高了后门攻击的鲁棒性和抗检测性。实验结果表明,在 MNIST 和 CIFAR-10 数据集上的攻击成功率均达到了 100%,且在模型微调和重训练后依然保持高效的攻击能力。与传统方法相比,本文在应对参数更新和防御净化操作时展示出更强的鲁棒性,并在多种检测方法中成功逃避了检测。这一创新设计不仅证明了结构后门在实际应用中的潜在威胁,更为后门攻击技术的发展提供了新的思路。未来可以进一步优化触发器的设计,并探索更复杂的攻击场景,以提升后门攻击的实用性和威胁性。

## 参考文献

- [1] LAURIOLA I, LAVELLI A, AIOLLI F. An introduction to deep learning in natural language processing: Models, techniques, and tools[J]. *Neurocomputing*, 2022, 470: 443-456.
- [2] MIN B, ROSS H, SULEM E, et al. Recent advances in natural language processing via large pre-trained language models: A survey[J]. *ACM Computing Surveys*, 2023, 56(2): 1-40.
- [3] ZAHRA A, PERWAIZ N, SHAHZAD M, et al. Person re-identification: A retrospective on domain specific open challenges and future trends[J]. *Pattern Recognition*, 2023, 142: 109669.
- [4] CHIB P S, SINGH P. Recent advancements in end-to-end autonomous driving using deep learning: A survey[J]. *IEEE Transactions on Intelligent Vehicles*, 2023, 9(1): 103-118.
- [5] MENGARA O, AVILA A, FALK T H. Backdoor Attacks to Deep Neural Networks: A Survey of the Literature, Challenges, and Future Research Directions [J]. *IEEE Access*, 2024, 12: 29004-29023.
- [6] LI Y, ZHANG S, WANG W, et al. Backdoor attacks to deep learning models and countermeasures: A survey[J]. *IEEE Open Journal of the Computer Society*, 2023, 4: 134-146.
- [7] LI Y, JIANG Y, LI Z, et al. Backdoor learning: A survey[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2022, 35(1): 5-22.
- [8] GUO W, TONDI B, BARNI M. An overview of backdoor attacks against deep neural networks and possible defences[J]. *IEEE Open Journal of Signal Processing*, 2022, 3: 261-287.
- [9] G U T, DOLAN-GAVITT B, GARG S. Badnets: Identifying vulnerabilities in the machine learning model supply chain[J]. *arXiv:1708.06733*, 2017.
- [10] CHEN X, LIU C, LI B, et al. Targeted backdoor attacks on deep learning systems using data poisoning[J]. *arXiv:1712.05526*, 2017.
- [11] LI S, XUE M, ZHAO B Z H, et al. Invisible backdoor attacks on deep neural networks via steganography and regularization[J]. *IEEE Transactions on Dependable and Secure Computing*, 2020, 18(5): 2088-2105.
- [12] CHENG S, TAO G, LIU Y, et al. Lotus: Evasive and resilient backdoor attacks through sub-partitioning[C]// *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024: 24798-24809.
- [13] HUANG Y, XU J F, GUO Q, et al. Personalization as a shortcut for few-shot backdoor attack against text-to-image diffusion models[C]// *Proceedings of the AAAI Conference on Artificial Intelligence*. 2024: 21169-21178.
- [14] NGUYEN T A, TRAN A. Input-aware dynamic backdoor attack [J]. *Advances in Neural Information Processing Systems*, 2020, 33: 3454-3464.
- [15] ZOU M, SHI Y, WANG C, et al. Potrojan: powerful neural-level trojan designs in deep learning models[J]. *arXiv:1802.03043*, 2018.
- [16] QI X, XIE T, PAN R, et al. Towards practical deployment-stage backdoor attack on deep neural networks[C]// *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022: 13347-13357.
- [17] BOBER-IRIZAR M, SHUMAILOV I, ZHAO Y, et al. Architectural backdoors in neural networks[C]// *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023: 24595-24604.
- [18] CLIFFORD E, SHUMAILOV I, ZHAO Y, et al. ImpNet: Imperceptible and blackbox-undetectable backdoors in compiled neural networks[C]// *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. IEEE, 2024: 344-357.
- [19] GAO Y, XU C, WANG D, et al. Strip: A defence against trojan attacks on deep neural networks[C]// *Proceedings of the 35th Annual Computer Security Applications Conference*. 2019: 113-125.
- [20] DOAN B G, ABBASNEJAD E, RANASINGHE D C. Februs: Input purification defense against trojan attacks on deep neural network systems[C]// *Proceedings of the 36th Annual Computer Security Applications Conference*. 2020: 897-912.
- [21] WANG B, YAO Y, SHAN S, et al. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks[C]// *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2019: 707-723.
- [22] ZHENG R, TANG R, LI J, et al. Pre-activation distributions expose backdoor neurons[J]. *Advances in Neural Information Processing Systems*, 2022, 35: 18667-18680.
- [23] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. *arXiv:1409.1556*, 2014.
- [24] LECUN Y. The MNIST database of handwritten digits [ EB / OL]. <http://yann.lecun.com/exdb/mnist/>.
- [25] KRIZHEVSKY A, HINTON G. Learning multiple layers of features from tiny images: TR-2009[R]. 2009.



**CHEN Xianyi**, born in 1986, Ph.D, associate professor, master's supervisor, is a member of CCF (No. 56536M). His main research interests include artificial intelligence security and big data security.



**CUI Qi**, born in 1994, Ph.D, associate professor, master's supervisor. His main research interests include information hiding and deep learning model security.