基于参数空间定向对抗扰动的后门检测与防御方法

田有亮[1,2,3]，金昆龙[1,2]，石璐嘉[1,2]，王帅[1,2]，左建烁[1,2]，向阿新[2,3]
（1. 贵州大学计算机科学与技术学院，贵州 贵阳 550025；2. 贵州省密码学与区块链技术特色重点实验室，贵州 贵阳 550025；
3. 贵州大学大数据与信息工程学院，贵州 贵阳 550025）

摘　要：为解决现有后门防御方法对显著且可分后门特征的依赖，以及触发器反演开销较高的问题，提出了参数空间定向对抗扰动框架 PTAP。该框架在参数空间内针对各候选目标类别，求解达到预设成功率所需的最小参数扰动幅度，并以该幅度作为后门异常检测的统计量，避免高开销的触发反演过程并提升检测性能。此外，PTAP 利用参数扰动指向的异常敏感方向来指导轻量级微调，在尽量保持主任务性能的同时削弱后门效应，并面向第三方模型场景实现检测与修复的一体化流程。在涵盖输入空间、特征空间和动态触发设置的 11 种后门攻击上的实验表明，PTAP 对后门目标的检测置信度超过 99%，显著降低了检测开销，并在各种攻击类型中保持稳定的性能。

关键词：深度神经网络；机器学习即服务；后门攻击；后门检测
中图分类号：TN92
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2026076

Backdoor detection and defense method via parameter-space targeted adversarial perturbations

Tian Youliang[1,2,3], Jin Kunlong[1,2], Shi Lujia[1,2], Wang Shuai[1,2], Zuo Jianshuo[1,2], Xiang Axin[2,3]
1. College of Computer Science and Technology, Guizhou University, Guiyang 550025, China
2. Guizhou Provincial Key Laboratory of Cryptography and Blockchain Technology, Guiyang 550025, China
3. College of Big Data and Information Engineering, Guizhou University, Guiyang 550025, China

Abstract: To address the reliance of existing backdoor defenses on salient and separable backdoor features, as well as their high trigger inversion cost, a parameter-space targeted adversarial perturbation (PTAP) framework was proposed. For each candidate target class in the parameter space, the minimum parameter perturbation was required to achieve a predefined success rate, and this quantity was used as the test statistic for backdoor anomaly detection, thereby avoiding costly trigger inversion and improving detection performance. Moreover, PTAP exploited the abnormally sensitive directions revealed by parameter perturbations to guide lightweight fine-tuning, thereby mitigating backdoor effects while largely preserving primary task performance and enabling an integrated detection-and-repair pipeline for third-party model scenarios. Experiments on eleven backdoor attacks covering input-space, feature-space, and dynamic-trigger settings show that PTAP achieves over 99% detection confidence for backdoor targets, significantly reduces detection overhead, and maintains stable performance across diverse attack types.

Keywords: deep neural network, machine learning as a service, backdoor attack, backdoor detection

收稿日期：2026–01–30；**修回日期：**2026–03–12
通信作者：金昆龙，kljin755@163.com
基金项目：国家重点研发计划基金资助项目（No.2025YFB3109800）；国家自然科学基金资助项目（No.62272123）；贵州省科技创新平台科研基金资助项目（No.CXPTXM[2025]024）；贵州省科技计划基金资助项目（No.[2020]5017, No.[2022]065）
Foundation Items: The National Key Research and Development Program of China (No.2025YFB3109800), The National Natural Science Foundation of China (No.62272123), Scientific and Technological Innovation Platform Research Project of Guizhou Province (No.CXPTXM[2025]024), Science and Technology Program of Guizhou Province (No.[2020]5017, No.[2022]065)

0 引言

深度神经网络（deep neural network, DNN）已广泛应用于图像分类[1]、人脸识别[2]和自动驾驶[3]等任务，并在实际系统中承担着关键功能。然而，训练高性能DNN往往需要大规模高质量数据、充足的计算资源以及烦琐的参数调优过程。同时，模型结构设计也具有较强的经验性，依赖于长期积累的工程与领域知识。对多数普通使用方而言，独立完成从数据准备到模型训练的过程不仅成本高、周期长，也常常难以落地。因此，出于对成本与交付周期的考虑，许多机构选择将模型训练外包给第三方机器学习即服务（machine learning as a service, MLaaS）[4]或直接复用在线模型库中的预训练模型，如 Caffe Model Zoo[5] 或 TensorFlow Model Zoo[6]。然而，这种外部依赖引入了新的安全风险[7]。攻击者可能在训练阶段通过篡改数据或在训练过程中植入后门，使模型在日常输入下表现正常，却在攻击者预设的触发条件出现时输出特定的错误预测。当该模型被集成到面向用户的服务中时，后门的影响可能扩散到更广泛的应用场景，进而造成严重后果。例如，在基于人脸识别的身份认证中，后门可能使冒名者在预设的触发条件下被误判为合法员工，从而绕过访问控制[8]。在自动驾驶等安全关键系统中，类似的触发条件甚至可能引发人身安全事故[9]。

目前，针对后门威胁已提出多类防御思路，其中在第三方模型复用场景中，触发反演被普遍视为后门检测与缓解的重要组成模块。该类方法通常以少量干净样本为基础，尝试恢复能够诱导模型输出目标类别的触发模式，典型代表是NC（neural cleanse）[10]、Tabor[11]、USB[12]等。它们在输入空间上对潜在触发器进行优化求解，并通过比较不同类别反演结果的各类统计异常来推断后门目标标签。然而，输入空间反演方法主要存在两方面局限：一是对于特征空间后门等攻击形态往往不够有效，反演得到的触发模式可能难以真实反映实际使用的触发模式；二是由于防御者缺乏目标类别信息，通常需要对全部类别进行逐一扫描与对比，计算开销较高。为弥补输入空间反演的不足，研究者进一步探索基于特征空间的检测与反演策略。例如，FeaTure-RE[13]通过分析带触发样本与干净样本在特征空间中的可分性来识别后门。BTI-DBF[14]则通过解耦良性特征，并在解耦空间中约束干净样本与其合成投毒样本的特征差异，以实现触发模式的恢复。总体而言，特征空间方法在覆盖攻击类型方面更具通用性，但近期研究[15]指出，特征空间方法往往依赖显著且可分的后门特征，在BadNets[16]、Blend[7]等简单触发场景中特征不够突出，易出现泛化缺陷并降低方法判别能力。

本文从一个新的角度探讨后门防御。不同于之前在输入与特征空间通过扰动样本反演触发器的方法，本文将扰动对象转移到参数空间，对模型参数施加定向对抗扰动，实现后门检测与缓解。触发反演方法的核心依赖于后门训练植入的一条捷径，该捷径更容易将部分输入推向后门目标类。本文发现，这种捷径不仅体现在输入空间，也体现在参数空间结构中。目标类附近的决策边界对某些参数方向呈现异常脆弱性。因此，只需对参数做轻微和定向的边界移动，就可能显著增强样本被判定为目标类的倾向。基于此，本文将能够最大化目标类输出的最小参数扰动定义为参数空间的定向对抗扰动（parameter-space targeted adversarial perturbation, PTAP），并以各候选类别对应扰动的统计异常来推断后门目标标签，从而检测后门模型。由于不依赖输入空间逐类扫描与触发器优化，该方法避免了输入空间反演在形态多样和计算耗时上的瓶颈，因此检测更高效。同时，无论后门采用特征空间机制、动态触发或者输入空间触发模式，后门训练都会在参数层面留下可被定向扰动放大的边界脆弱性，因此参数空间的定向对抗扰动能够捕获这种特性并增强对不同种类后门攻击的覆盖能力。进一步地，本文利用该扰动指向的敏感方向指导微调过程，在尽量保持主任务性能的同时削弱后门效应，实现模型修复。

本文的主要工作如下。

1) 提出了一种基于参数空间定向对抗扰动的后门检测方法。该方法不依赖触发反演，而是通过判别模型参数空间中是否存在异常显著的定向对抗扰动来识别后门目标标签，从而避免了反演过程中较高的开销，并缓解特征空间方法在弱表征场景下检测不稳定的问题。

2) 提出了定向对抗扰动不仅可用于模型检测，还可用于指示后门在参数空间的敏感方向，从而为模型修复提供指引。本文沿抑制该异常敏感性的方向进行轻量微调，在尽量保持主任务性能的同时削弱后门效应。由此，本文形成了一个面向MLaaS场

景的检测与缓解一体化框架，使检测完成后不需要高成本重训即可实现后门模型快速修复。

3) 在多种攻击与防御设置下进行了对比实验验证。实验覆盖 11 类代表性后门攻击与 8 种典型后门防御方法，结果表明，PTAP 在显著降低计算开销的同时，对输入空间后门与特征空间后门均表现出超过 99% 的检测置信度，并能够覆盖多种攻击形态。

1　相关工作

1.1　后门攻击

根据触发模式的注入方式，后门攻击分为两类：一类攻击以训练数据投毒为手段，在输入空间对少量样本进行修改并将其标签指向预设目标类别，从而使模型在触发条件出现时输出攻击者指定结果；另一类攻击则在更强威胁假设下发生，攻击者能够干预训练过程或直接改写模型参数，使后门以更隐蔽的形式嵌入特征空间中。

在输入空间型攻击中，触发器的设计决定了其可见性与泛化特性。按触发模式是否固定，可进一步分为静态与动态触发两种范式。静态触发在所有被投毒样本上复用同一模式，如固定形状的补丁[16]、噪声叠加[7]或特定的对抗性扰动[17]，其中 BadNets[16] 是早期典型代表之一。Blend[7]则通过将半透明图像与输入进行混合来实现触发注入。与之相对，动态触发更强调样本相关性与隐蔽性，触发器会随样本变化而变化，从而降低可被反演与检测的风险。例如，SSBA（sample special backdoor attack）[18]利用隐写式编码机制为每个样本生成特定触发模式，在几乎不改变视觉外观的情况下诱导模型在触发时输出目标标签。IAD（input-aware dynamic）[19]则显式地构造了输入相关的触发器，并通过交叉触发测试衡量触发的可复用性。WaNet[20]采用平滑的几何翘曲作为触发信号，使输入变化难以被人眼察觉且具有良好的触发一致性。BppAttack[21]进一步利用人类视觉感知的弱点，通过量化与抖动等操作嵌入更不易被察觉的触发扰动，在几乎不改变视觉外观的情况下诱导模型在触发时输出目标标签。

特征空间相关的后门攻击通常假设攻击者具备更高权限，能够操纵训练目标、优化流程或直接篡改模型权重，从而绕开仅依赖输入空间统计规律的防御。此类攻击既可以通过在训练过程中引入特定优化约束来嵌入后门特征，也可以通过修改权重使模型在内部表示层面形成对触发条件的异常敏感性。例如，ITI（invisible trigger image）[22]通过将触发图像的隐藏特征嵌入目标图像中，生成有效且隐形的触发条件。DFST（deep feature space Trojan）[23]通过操控深层特征将后门效应更多地固化在特征空间中，使输入层触发模式不易被直接反演，并对多类防御方法形成挑战。DEFEAT[24]用自适应扰动投毒，同时加入潜在特征约束，使触发样本在隐层表征上更像目标类，从而绕过基于特征异常的检测。近年来，自适应策略与动态特征空间触发器的结合在后门攻击中得到了广泛应用，以提升后门攻击的隐蔽性与有效性。PBADT（precise backdoor attack with dynamic trigger）[25]通过自适应选择触发位置，并结合特征激活反馈信息，精确地施加触发器，从而增强了攻击的精确度和隐蔽性。文献[26]则通过动态频域隐写触发器（dynamic frequency domain trigger, DFDT）策略，保证触发器对不同输入样本的适应性。本文将覆盖各种类型的攻击，包括输入空间、特征空间、动态触发和自适应攻击，对 PTAP 进行测试，以提供一个类似于实际威胁的系统评估。

1.2　后门防御

后门防御通常包含后门检测与后门缓解两种基本方法。后门检测方法的目标是识别被植入后门的模型或样本。面向后门模型检测通过分析模型行为或统计特征判定其是否异常[27]。面向后门样本检测在推理阶段对输入进行判别与过滤[28-29]。此外，部分检测方法进一步采用反向工程思想，对可疑模型进行触发器恢复，通过触发器的异常统计特征定位目标标签并提升检测可信度[30-31]。相比之下，后门缓解方法的目标是在尽量保持主任务性能的前提下，从已感染模型中削弱或移除后门效应。现有技术路线包括基于干净数据的微调、知识蒸馏[32]、遗忘学习[33]、剪枝[34]以及训练阶段防御[35-36]等。值得注意的是，触发器反演在多类后门检测与缓解框架中扮演关键模块，其核心是从模型行为出发反演触发模式。按反演空间不同，大体可分为像素空间反演与特征空间反演两类。像素空间反演以 NC[10]为代表，通过在输入域优化通用扰动及其掩码来反推潜在触发器，随后大量工作围绕优化稳定性与计算效率进行改进，如取消掩码约束的优化策略[30]、通过多候选触发集合减少迭代开销[31]、引入选择性优化等策略[37]降低搜索成本。近年来，

研究重心逐步转向特征空间约束，FeaTure-RE[13]利用后门相关激活在表示空间中的结构性特征进行约束与判别。UNICORN[38]则通过构造从输入空间到其他空间的映射，将检测信号投影到更易分离的表示域以增强鲁棒性。除通用分类场景外，后门防御也在特定任务场景中得到扩展。例如，面向目标检测的 TTBD（test-time backdoor detection）[39]侧重推理阶段的后门输入检测，通过语义感知变换下的预测一致性差异实现区分。NaviDet[40]则面向图像条件扩散模型，针对生成模型的后门风险提供检测与防护机制。本文仅关注通用后门检测与后门缓解两类防御设置，面向 MLaaS 场景，在该场景下防御者通常只能获得可疑模型和少量本地干净样本。因此，本文侧重于在缺乏攻击先验的条件下，提供可操作的模型检测和修复策略。

2 预备知识

本节提供支撑本文方法设计的预备知识，包括 DNN 后门攻击的形式化定义、触发模式的构造方式以及相应的训练与防御设定。通过建立统一的表示框架，本节为后续系统模型与 PTAP 的算法描述建立统一的技术语境。本文涉及的主要符号及其含义如表 1 所示。

表 1 符号说明

符号	说明
$\mathcal{S}$	后门触发器模式
$\mathcal{M}$	后门触发器掩码
$\mathcal{D}$	训练数据集
K	数据集中的标签数量
N	待检的 DNN 模型
θ	模型参数集合
$\mathcal{X}$	训练样本
$\mathcal{X}^*$	植入触发器的后门样本
$\mathcal{Y}$	正常样本标签
ζ	后门目标标签
λ	后门缓解目标函数的权重因子
δ	输入样本的扰动
$\Delta\theta$	参数的扰动
$\mathcal{L}$	交叉熵损失函数
ρ	错误分类的成功率阈值
$\mathcal{R}$	参数定向扰动强度集合
$\mathcal{T}$	可疑的后门目标标签集合

2.1 后门攻击概述

本文将 DNN 中的后门攻击定义为一种仅在输入包含特定触发器 $\mathcal{S}$ 时才被激活的隐藏行为模式。设标准 K 类分类任务的训练集为 $\mathcal{D}=\{(x_i,y_i)\}_{i=1}^{K}\subset\mathcal{X}\times\mathcal{Y}$，其中 $\mathcal{X}\subset\mathbb{R}^d$，$\mathcal{Y}=\{1,\cdots,K\}$。对于图像任务，触发器由局部模式 $\mathcal{S}$ 及其掩码 $\mathcal{M}$ 描述，掩码 $\mathcal{M}$ 和目标样本 $\mathcal{X}$ 是相互同构的矩阵。当掩码 $\mathcal{M}$ 对应位置的元素为 0 时，保留目标样本 $\mathcal{X}$ 中的原始像素；反之，将掩码对应位置的像素与局部触发模式 $\mathcal{S}$ 按与掩码成比例融合。因此，后门攻击可以看作在干净子集 $\mathcal{D}_{\mathrm{c}}=\mathcal{D}-\mathcal{D}_{\mathrm{b}}$ 和后门子集 $\mathcal{D}_{\mathrm{b}}$ 上的多任务学习问题。具体来说，攻击样本及其标签 $(\mathcal{X}^*,\zeta)\in\mathcal{D}_{\mathrm{b}}$ 可以通过规则 $\mathcal{F}_{\mathcal{M}}$ 将触发模式整合到良性目标样本 $\mathcal{X}$ 中来生成。

$$\mathcal{X}^*=\mathcal{F}_{\mathcal{M},\mathcal{S}}(\mathcal{X})=\mathcal{M}\odot\mathcal{S}+(1-\mathcal{M})\odot\mathcal{X} \tag{1}$$

其中，$\odot$ 表示逐元素乘。攻击者将部分样本替换为 $(\mathcal{X}^*,\zeta)\in\mathcal{D}_{\mathrm{b}}$，并与干净样本共同训练，诱导模型 N 同时满足

$$\arg\min_{\theta}\left[\underbrace{\mathbb{E}_{(x,y)\in\mathcal{D}_{\mathrm{c}}}\mathcal{L}(N(x;\theta),y)}_{\text{干净样本}}+\underbrace{\mathbb{E}_{(x,y)\in\mathcal{D}_{\mathrm{b}}}\mathcal{L}(N(\mathcal{F}_{\mathcal{M},\mathcal{S}}(x);\theta),\zeta)}_{\text{后门样本}}\right] \tag{2}$$

其中，$\mathcal{L}$ 表示损失函数，θ 表示模型 N 的参数空间，$\mathbb{E}$ 表示期望。在常规后门攻击中，触发特征与任务相关特征通常彼此独立[41]，因此模型在干净样本上能够保持正常准确率，而在触发输入上呈现高成功率（success rate, SR）。训练过程不断迭代上述混合数据的学习，直到模型同时拟合主任务与后门映射。最终，攻击者可通过向发布模型的任意输入叠加 $\mathcal{F}_{\mathcal{M},\mathcal{S}}(x)$ 来激活后门行为。

2.2 后门防御概述

为保证上线模型的安全性并防范潜在后门攻击，服务商通常在发布前部署后门防御算法。典型地，给定待检的 DNN 模型 N 以及一小组干净子集 $\mathcal{D}_{\mathrm{c}}=\{(x_i,y_i)\}_{i=1}^{n}$（$n$ 表示样本数量），防御者通过求解式(3)所示优化问题来恢复潜在触发器模式。

$$\arg\min_{\mathcal{M},\mathcal{S}}\mathcal{L}(N(\mathcal{F}_{\mathcal{M},\mathcal{S}}(x);\theta),\zeta)+\lambda|\mathcal{M}| \tag{3}$$

其中，λ 为正则权重。大量现有工作[10,30-31]基于这一范式，或者以此为基础，从输入空间转向特征空

间约束对所有可能的目标标签进行触发反演。同时，后门缓解方法在后门防御中也是必要的。虽然在检测到后门后，服务商可以选择拒绝该模型并另行获取替代模型或训练服务，但这在实际应用中往往不可行。首先，重新训练通常需要大量资源与专业技术支持，找到合适的训练服务本身就可能具有挑战性。例如，服务商可能受限于特定教师模型的可用性，或面临某些替代模型无法支持特定任务。其次，服务商在许多场景下只能访问受感染模型及少量验证数据，而无法获得原始训练数据。在此情况下，重新训练不可实施，后门缓解成为唯一可行的选择，而这正是 PTAP 关注的防御场景。

3　系统模型

本节给出了 PTAP 的系统模型。如图 1 所示，PTAP 系统架构由两个协同模块构成：PTAP-Detect 负责后门检测；PTAP-Project 负责后门缓解与修复。PTAP 主要部署在 MLaaS 平台侧，面向具备模型权重访问权限的白盒防御场景，用于在模型上线前对后门风险进行评估与处理。因此，在系统模型中，PTAP 部署于 MLaaS 平台的模型接入层，作为模型安全控制组件，对来自第三方训练、外包学习或联邦学习节点的模型执行上线前的后门防御处理，确保纳入服务目录的模型满足基本后门安全性要求，从而为用户提供可信的推理服务与安全保障。

3.1　核心对象

基于 PTAP 系统架构，系统模型由以下 3 个实体组成。

1) 云服务提供方（cloud service provider, CSP）。在 PTAP 系统架构中，CSP 是负责模型接入与上线审核的主体，也是后门防御的核心责任方，在系统模型中执行模型部署任务。CSP 具有对模型的白盒访问权限，需要防御来自第三方训练、外包学习或联邦学习恶意参与者节点的模型投毒与后门注入攻击，确保用户仅访问经过安全验证的模型服务。

2) 用户 $\mathcal{C}$。用户 $\mathcal{C}$ 是在系统模型中通过 MLaaS 接口调用云端模型服务的各类终端用户或组织，仅具备对模型输入与输出结果的访问权限，无权查看模型结构、参数和训练数据。用户可以在应用层实施有限的输入与输出异常检测，但无法对云端模型本身进行分析或修复，因此在后门防御中安全能力受限，需依赖 CSP 提供的模型安全保障。

3) 模型供应方 $\mathcal{P}$。模型供应方 $\mathcal{P}$ 指向云平台提交预训练模型的第三方模型实体。当 CSP 自研模型时，$\mathcal{P}$ 与 CSP 为同一主体。在外包训练、第三方模型市场、教师模型或联邦学习等场景中，$\mathcal{P}$ 与 CSP 相互独立，构成第三方模型来源。在实际攻击场景中，$\mathcal{P}$ 既可能因训练数据受污染而在不知情的情况下生成带后门模型，也可能作为主动攻击者有意植入后门。因此，$\mathcal{P}$ 处于模型安全的上游，其行为将直接影响平台侧 CSP 的模型安全性以及下游用户 $\mathcal{C}$ 所面临的风险。

3.2　威胁模型

本文采用与现有研究一致的威胁模型[10,13,38]。攻击者直接向防御者提供后门模型 $\hat{\mathcal{N}}$，该模型在主任务上性能表现与良性模型相近，但内部植入后门。后门可由触发函数 $\mathcal{F}_{\mathcal{M},\mathcal{S}}(x)$ 激活，使模型以高

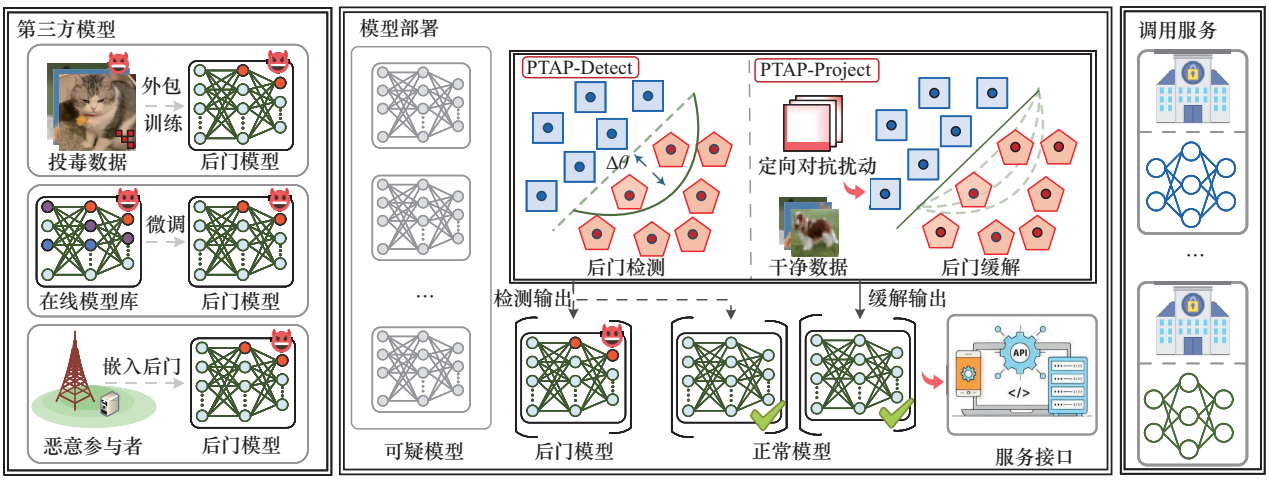


图 1　PTAP 系统架构

攻击成功率（attack success rate, ASR）输出攻击者指定的目标标签ζ。攻击者无法控制模型在平台侧的后续处理与部署方式，但可在模型上线后通过向系统输入$\mathcal{F}_{\mathcal{M},\mathcal{S}}(x)$的方式利用该后门。

1) 防御者能力。防御者接收到模型$\tilde{\mathcal{N}}$后，并不了解其训练过程、训练数据或是否带有后门，也不知道潜在触发器的模式。与已有工作一致[30,37]，假设防御者可以访问训练好的DNN和一组正确标记的样本来测试模型性能，还可以访问计算资源以测试或修改DNN，如GPU或基于GPU的云服务。

2) 防御者目标。防御者的最终目标是在模型部署期间抑制ASR，同时保留尽可能高的干净任务准确率（clean accuracy, CA）。首先，防御者需要判断给定模型是否已被后门篡改，并在检测到异常时识别其目标标签。其次，防御措施的目标是使后门失效，同时尽量保留模型在干净样本上的正常性能。

4 方案设计

为实现第3节提出的安全目标，本节详细介绍了PTAP的两个核心模块PTAP-Detect与PTAP-Project的设计。PTAP-Detect通过在参数空间中刻画模型对不同目标标签的敏感性，判断模型是否存在后门；PTAP-Project在此基础上对模型参数进行受控微调，在尽量保持主任务精度的前提下削弱或清除已识别的后门行为。

4.1 防御直觉和概述

本文从基于触发反演的后门防御[10,12]中得到技术背后的直觉。如图2所示，触发反演的后门防御将分类问题视为在多维空间中创建分区，每个维度捕获一些特征。后门触发器训练创建从属于其他类标签的空间区域到属于A的区域的捷径，感染模型显示了沿触发器维度的新边界，因此标签B或C中的任何输入都可以通过移动一小段距离而被错误分类为标签A。令$\mathcal{Y}$表示DNN模型中输出标签的集合。考虑标签$\mathcal{Y}_i \in \mathcal{Y}$和攻击的目标标签$\zeta$，如果存在将分类为$\zeta$的触发输入$\mathcal{F}_{\mathcal{M},\mathcal{S}}(x)$，则将$\mathcal{Y}_i$的所有输入转换为$\zeta$所需的最小输入扰动量，且受触发器大小的限制，即$\delta_{\forall\to\zeta} \leqslant \left|\mathcal{F}_{\mathcal{M},\mathcal{S}}(x)\right|$，其中$\delta_{\forall\to\zeta}$表示将任何输入分类为$\zeta$所需的最小输入扰动量。因此，如果后门触发器$\mathcal{F}_{\mathcal{M},\mathcal{S}}(x)$存在，则有

$$\delta_{\forall\to\zeta} \leqslant \left|\mathcal{F}_{\mathcal{M},\mathcal{S}}(x)\right| \ll \min_{i,i\neq\zeta}\delta_{\forall\to i} \tag{4}$$

触发反演的后门防御通过后门检测所有输出标签中$\delta_{\forall\to i}$的异常低值来检测触发器$\mathcal{F}_{\mathcal{M},\mathcal{S}}(x)$。受此启发，这种捷径不仅存在于输入空间，参数空间也存在类似的情况。如图2(b)所示，由于后门感染模型触发器维度的新边界存在，对该边界进行定向对抗扰动，将使原本难以分类成标签A的C类样本通过少量轻微的边界偏移被分类成标签A。

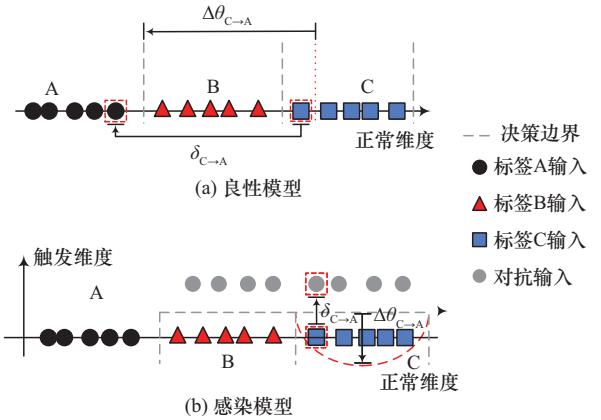


图2 后门触发导致的输入空间与参数空间边界偏移

观察给定K分类后门感染模型$\tilde{N}$，模型输出$f(x;\theta)\in\mathbb{R}^K$，固定目标类$t$。定义其他标签与目标标签的间隔为

$$m_t(x;\theta) = f_t(x;\theta) - \max_{k\neq t} f_k(x;\theta) \tag{5}$$

即x被分类为目标类t等价于$m_t(x,\theta)>0$。由于max函数在竞争类切换处不可微，故用LSE（log-sum-exp）替代max函数，则分类间隔可定义为

$$\tilde{m}_t(x;\theta) = f_t(x;\theta) - \tau\ln\sum_{k\neq t}\exp\left(\frac{f_k(x;\theta)}{\tau}\right) \tag{6}$$

其中，$\tau>0$，当$\tau\to 0$，$\tilde{m}_t$一致逼近m_t。考虑沿方向$v\neq 0$的参数扰动$\theta(\alpha)=\theta+\alpha v$，其中$\alpha$是缩放系数。由于$\tilde{m}_t$可微，在$\alpha=0$处做一阶泰勒展开有

$$\tilde{m}_t(x;\theta+\alpha v) = \tilde{m}_t(x;\theta) + \alpha\left\langle\nabla_\theta\tilde{m}_t(x;\theta),v\right\rangle + o(\alpha) \tag{7}$$

令$p_t(x;v)=\left\langle\nabla_\theta\tilde{m}_t(x;\theta),v\right\rangle$，则在局部可近似写为

$$\tilde{m}_t(x;\theta+\alpha v) \approx \tilde{m}_t(x;\theta) + \alpha p_t(x;v) \tag{8}$$

因此，若$\tilde{m}_t(x;\theta)>0$，则不需要优化；若$\tilde{m}_t(x;\theta)<0$且$p_t(x;v)<0$，则一阶近似下沿该方向无法推动其跨界；若$\tilde{m}_t(x;\theta)<0$且$p_t(x;v)>0$，则跨界所需的最小步长在一阶近似下有

$$\beta_t(x;v)=\frac{-\tilde{m}_t(x;\theta)}{p_t(x;v)} \tag{9}$$

后门检测与定向对抗扰动关注的是让一批非目标样本中至少ρ比例被诱导为t，因此在非目标样本集$\mathcal{D}_{\neg t}=\{(x,y)|y\neq t\}$中定义SR为

$$\mathrm{SR}_t(\alpha,v)=\frac{1}{\mathcal{D}_{\neg t}}\sum_{x\in D_{\neg t}}\mathbb{I}\left(\tilde{N}(x;\theta+\alpha v)=t\right) \tag{10}$$

其中，$\mathbb{I}$是指示函数，在理想情况下样本被诱导为t等价于α不小于该样本跨界步长$\beta_t(x;v)$，因此在一阶近似下有

$$\mathbb{I}\left(\tilde{N}(x;\theta+\alpha v)=t\right)\approx\mathbb{I}\left(\alpha\geqslant\beta_t(x;v)\right) \tag{11}$$

代入经验成功率可得

$$\mathrm{SR}_t(\alpha;v)=\frac{1}{\mathcal{D}_{\neg t}}\sum_{x\in D_{\neg t}}\mathbb{I}\left(\alpha\geqslant\beta_t(x;v)\right) \tag{12}$$

因此，至少ρ比例样本被诱导为t的最小阈值$\beta_\rho(t;v)$可解释为

$$\beta_\rho(t;v)=\inf\left\{\alpha\geqslant 0:\mathrm{SR}_t(\alpha;v)\geqslant\rho\right\} \tag{13}$$

在参数空间中，采用ℓ_1范数衡量扰动代价，则沿方向v的扰动$\Delta\theta=\alpha v$的ℓ_1代价为$\|\Delta\theta\|=\alpha\|v\|$，定义达到$\rho$比例的最小$\ell_1$代价为

$$\beta_{\rho,\ell_1}(t;v)=\inf\left\{\|\Delta\theta\|:\Delta\theta=\alpha v,\mathrm{SR}_t(\alpha;v)\geqslant\rho\right\}=\|v\|\beta_\rho(t;v) \tag{14}$$

其中，$\beta_\rho(t;v)$表示沿方向v施加最小缩放α，即可使非目标集$\mathcal{D}_{\neg t}$中至少ρ比例样本跨越决策边界并被预测为t；$\beta_{\rho,\ell_1}(t;v)$为将该最小缩放换算为ℓ_1范数下的扰动代价。这两个量越小，意味着边界对该方向上的参数扰动越敏感。本文使用最原始的4×4后门白块在CIFAR10数据集中对标签0植入后门，对所有标签扫描计算上述指标，并通过多次随机重启与重采样检验其稳定性，观察后门目标类是否在这两个指标上呈现稳定的下尾离群。

图3给出了对所有候选标签扫描得到的50次$\beta_\rho(t;v)$和$\beta_{\rho,\ell_1}(t;v)$散点分布。结果显示，除目标标签外，其他标签的扰动代价集中在较大范围内，后门目标标签（label=0）在两项代价上同时显著偏小，形成稳定的下尾离群点。这提示后门训练在参数空间中为目标类引入了更易被统一方向推动的捷径，从而使达到同等成功率门槛所需的最小代价显著降低。

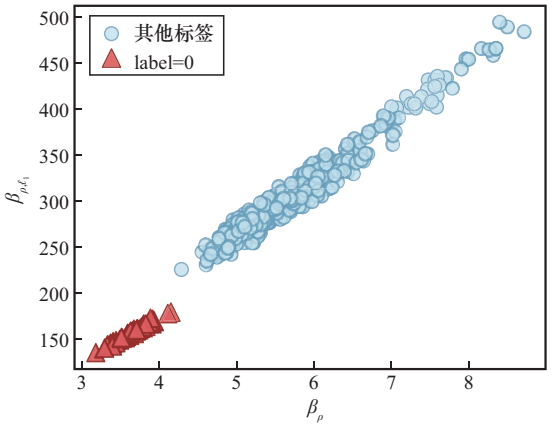


图3　各目标标签的扰动代价散点分布

基于以上观察，如果存在将分类为ζ的触发输入$\mathcal{F}_{\mathcal{M},\mathcal{S}}(x)$，则将$\mathcal{Y}_i$与$\zeta$进行分类的参数边界要比$\mathcal{Y}_i$与其他正常标签$\mathcal{Y}_{j,j\neq\{i,\zeta\}}\in\mathcal{Y}$更敏感，因此，如果后门触发输入$\mathcal{F}_{\mathcal{M},\mathcal{S}}(x)$存在，则有

$$\Delta\theta_{\forall\to\zeta}\ll\min_{i,i\neq\zeta}\Delta\theta_{\forall\to i} \tag{15}$$

其中，$\Delta\theta_{\forall\to\zeta}$表示将任何输入分类为$\zeta$所需的最小参数扰动量。本文将在后续实验中证明参数空间定向对抗扰动相较于输入空间扰动的先进性。

4.2　检测算法

本节详细介绍基于参数空间定向对抗扰动的后门检测方法PTAP-Detect，具体检测算法如算法1所示。

算法1　后门模型检测

输入　模型N，干净样本集$\mathcal{D}$，成功率约束ρ

输出　潜在后门目标标签$\mathcal{T}$

1) 初始化$\mathcal{R}\leftarrow\varnothing$，$\mathcal{T}\leftarrow\varnothing$
2) for t in $\{0,\cdots,K-1\}$ do
3) for $\mathcal{P}$ in $\{\mathcal{P}_1,\mathcal{P}_2,\cdots,\mathcal{P}_m\}$ do
4) 生成t的评估集$\mathcal{D}_{\neg t}\leftarrow\mathcal{D}$
5) 优化t的对抗扰动$\Delta\theta_t=\arg\min_{\Delta\theta}\sum_{x\in D'}\mathcal{L}\left(N(x;\theta+\Delta\theta),t\right)$，其中$D'\subseteq\mathcal{D}_{\neg t},|D'|\geqslant\rho|\mathcal{D}_{\neg t}|$
6) 更新扰动强度集合$\mathcal{R}\leftarrow\mathcal{R}\cup\{|\Delta\theta_t|\}$
7) end for
8) end for
9) for t in $\{0,\cdots,K-1\}$ do
10) if $\mathrm{MAD}(\mathcal{R},|\Delta\theta_t|)\geqslant 1.96$ then
11) $\mathcal{T}\leftarrow\mathcal{T}\cup\{t\}$
12) end if

13) end for

14) return $\mathcal{T}$

首先给出参数空间定向对抗扰动的优化求解过程，该过程作为后门检测流程的第一步，用于计算各目标标签对应的最小定向对抗扰动。

针对每一个候选目标类别$t \in \{0,\cdots,K-1\}$，算法在参数空间中构造一个定向攻击问题。假设由所有真实标签不等于t的样本构成的非目标样本的评估集为$\mathcal{D}_{\neg t}=\{(x,y)|y \neq t\}$，则PTAP-Detect希望通过仅优化扰动参数$\Delta\theta$，使$\mathcal{D}_{\neg t}$在带扰动模型$\mathbb{N}_{\Delta\theta}$下尽可能地被预测为标签$t$。形式上，对每个标签$t$需求解如式(16)所示的约束优化问题。

$$\arg\min_{\Delta\theta}\sum_{x\in\mathcal{D}_{\neg t}}\mathcal{L}(N(x;\theta+\Delta\theta),t) \tag{16}$$

本文通过随机梯度下降更新所有对抗扰动。为了将优化问题的难易程度刻画为可比较的标量，算法引入攻击成功率作为判据。对任意给定的扰动参数$\Delta\theta$，在评估集$\mathcal{D}_{\neg t}$上计算被分类为目标标签t的样本比例，当样本比例不小于预设的成功率约束ρ时，认为已经在参数空间中成功构造出目标标签t的定向扰动。将对应扰动向量的总扰动强度$\left|\Delta\theta_t\right|$作为实现将$\mathcal{D}_{\neg t}$分类为目标标签$t$所需的最小扰动代价，对所有目标类别重复上述过程，即可得到按目标类生成的扰动强度集合$\mathcal{R}=\{\left|\Delta\theta_t\right|\}_{t=0}^{K-1}$。

如4.1节所述，若模型存在后门，则针对真实的后门目标标签ζ，模型参数空间中已形成一条隐式捷径，仅需在参数空间施加极小扰动，即可将绝大多数非ζ类样本诱导至目标标签。这表明在相同成功率约束ρ下，后门目标标签对应的定向扰动$\left|\Delta\theta_{\forall\to\zeta}\right|$应显著小于正常类别对应的$\left|\Delta\theta_{\forall\to i,i\neq\zeta}\right|$。基于该观察，本文通过优化过程分别求解各目标标签的最小定向参数扰动，并得到其总扰动强度$\left|\Delta\theta_t\right|$。随后，在所有标签对应的扰动强度分布中，后门目标标签将表现为具有异常小的离群值。为鲁棒地识别此类低幅度扰动异常，PTAP-Detect采用基于中位数绝对偏差（median absolute deviation, MAD）的单侧异常检测方法。该方法在多个异常点时仍具有良好的稳健性[42]。首先计算各数据点相对于中位数的绝对偏差，该绝对偏差的中位数即MAD，用于刻画扰动强度分布的离散程度。随后，将每个数据点的异常指数定义为该绝对偏差与MAD的比值，并在假设基础分布近似服从正态分布的条件下，采用常数1.482 6对异常指数进行归一化。在此设置下，异常指数大于1.96的数据点可在95%的置信水平下被判定为异常值。本文将扰动强度集合$\mathcal{R}=\{|\Delta\theta|_t\}_{t=0}^{K-1}$中满足该条件的标签放入潜在后门目标标签$\mathcal{T}=\{\zeta_1',\zeta_2',\cdots\}$，并仅关注分布下尾处的离群点。

4.3 缓解算法

在4.2节中，PTAP已经通过参数空间定向扰动和MAD异常检测，得到一组疑似后门目标标签集合。本节在此基础上给出后门缓解的完整算法PTAP-Project的设计。

虽然PTAP-Detect已经生成一组参数对抗扰动，但为了避免大量标签存在时的运行开销，PTAP-Detect中ρ的值不宜取太大，故参数对抗扰动的粒度受限于成功率约束ρ。因此，PTAP-Project先对所有潜在后门目标标签ζ'求解更精细（$\rho>0.9$）的扰动解集合$\Delta'=\{\Delta\theta_1',\Delta\theta_2',\cdots\}$，$\Delta'$刻画了模型在参数空间中通往后门映射的方向。在后门缓解阶段，CSP掌握一个小规模干净数据集$\mathcal{D}_{\mathrm{c}}=\{(x_i,y_i)\}_{i=1}^{n}$，目标是在尽量保持干净样本精度的前提下削弱模型的后门效应。在PTAP中，PTAP-Project通过利用Δ'显式地构造后门扰动子空间，并约束微调的梯度更新不再沿该子空间方向移动，同时逐步减弱现有权重在该子空间上的投影，从而削弱模型后门效应。PTAP的缓解模块以定向扰动为核心，将检测阶段的结果转化为微调的几何约束。具体而言，PTAP-Project通过正交化构造Δ'的线性子空间。

$$U=\mathrm{span}\{u_1,u_2,\cdots,u_r\}\subset\mathbb{R}^d \tag{17}$$

其中，u_i为正交归一基向量，d为参数总维度，U为后门扰动子空间。直观地，U捕捉了模型在参数空间中导向被发现后门的主要变化模式，其正交补$U^{\perp}$则可理解为相对安全的更新方向集合。因此，在缓解阶段，PTAP不再无约束地更新模型，而是要求所有参数更新落在$U^{\perp}$中，这一约束等价于对每一步梯度沿U的分量进行投影抑制，仅保留其在$U^{\perp}$上的部分参数用于更新。为进一步削弱模型当前权重在后门子空间上已经形成的投影，PTAP额外引入方向正则项$\mathcal{L}_U$惩罚模型N在U上的分量，即

$$\mathcal{L}_U(\theta)=\sum_{u\in U}(\langle\theta,u\rangle)^2 \tag{18}$$

则微调过程被表述为带几何约束的优化问题，即

$$\arg\min_\theta \mathbb{E}_{(x,y)\in\mathcal{D}_c}\left[\mathcal{L}(N(x;\theta),y)+\lambda\mathcal{L}_U(\theta)\right] \tag{19}$$

其中，λ为较小的权重系数，用于平缓地惩罚模型在后门扰动子空间上的投影。PTAP-Project将定向扰动的参数几何信息贯穿于后门检测与缓解两个阶段。

5　实验分析

本节通过实验对本文方法的有效性进行评估，实验内容涵盖基准稳健性、消融研究以及有限计算资源条件下的性能分析。所有实验均在配备NVIDIA A40 GPU的计算环境中完成。

5.1　实验参数配置

1) 攻击设置

本文针对11类具有代表性和挑战性的后门攻击方法对PTAP的检测性能进行了评估，涵盖多种典型攻击范式，包括静态后门攻击BadNets[16]、Trojan[43]和Blend[7]，动态后门攻击Bpp[21]、SSBA[18]、ITI[22]和WaNet[20]，特征空间后门攻击DFST[23]、DEFEAT[24]与自适应后门攻击PBADT[25]、DFDT[26]，除非另有说明，大多数攻击均采用其原始论文或开源代码中的默认设置，包括后门触发器的模式与尺寸配置。所有攻击的目标标签统一设为类别0，默认投毒率φ为10%，相关攻击基于BackdoorBench框架[41]实现。实验分别在配置了ResNet18的CIFAR10与GTSRB数据集，以及配置了ResNet34的Tiny-ImageNet与ImageNet数据集（200类）上进行。模型训练过程中采用随机梯度下降优化器，初始学习率设为0.1，动量设为0.9，权重衰减系数设为5×10^{-4}。其中，CIFAR10数据集训练200轮，批次大小为128，ImageNet与Tiny-ImageNet数据集训练300轮。训练过程中使用余弦学习率调度策略对学习率进行调整。

2) 防御设置

本文将PTAP与8种后门防御方法进行了比较。这些方法包括5种后门缓解方法：Fine-pruning[44]、NC[10]、NAD（neural attention distillation）[32]、FeaTure-RE[13]和I-BAU[33]，以及5种后门检测方法：NC[10]、FeaTure-RE[13]、Tabor[11]、Pixel[30]和UNICORN[38]。所有后门防御方法在实验中均仅允许访问占训练集5%的干净样本，以模拟现实场景中防御方对良性数据的受限获取条件。各对比方法的超参数均依据其公开实现进行调节，并针对不同攻击场景选取性能最优的配置。对于PTAP，相关超参数设置如下：阈值参数ρ设为0.4，正则化系数λ设为0.01，微调轮数设为30，学习率设为0.01。

3) 评估指标

本文采用干净任务准确率CA和攻击成功率ASR评估后门缓解性能，其中CA表示模型在干净测试集上的分类准确率，ASR表示模型在后门测试集上的攻击成功率。此外，采用后门模型检测率（detection rate, DR）评估后门检测性能，其中DR表示检测方法正确识别后门模型的比例。

5.2　主要防御结果

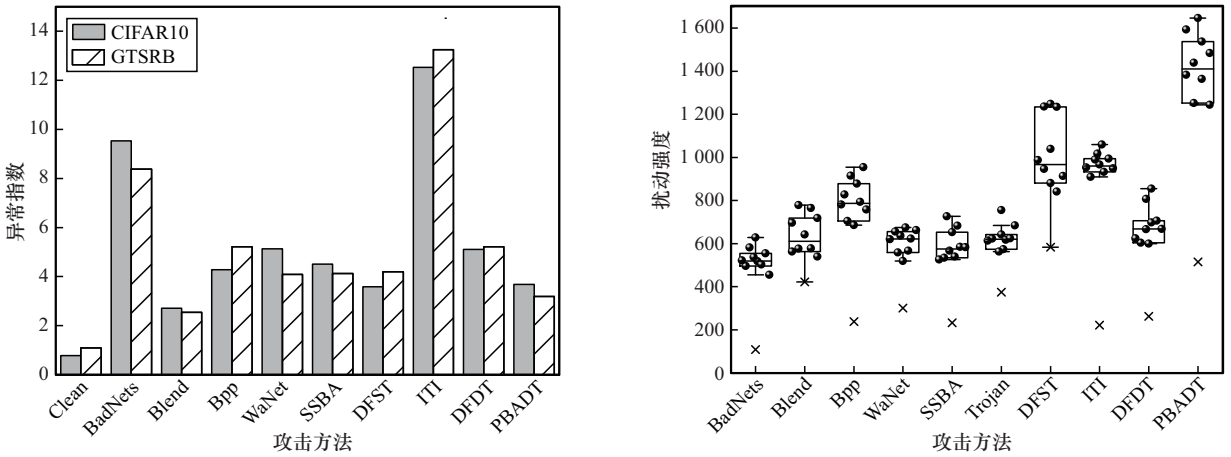
1) 检测性能

本文在不同攻击设置下训练并检测了20个后门模型，检测结果如表2所示，其中加粗数据表示该设置下的最高检测率。PTAP在CIFAR10与GTSRB数据集上均取得较高检测率，且在WaNet、Bpp、DFST、DFDT与PBADT等更具挑战性的攻击场景中优势更为明显。在BadNets、Blend等简单触发器攻击场景中，FeaTure-RE、UNICORN等后门防御方法相较NC出现小幅回落，本文认为其原因主要与攻击模式及方法偏好有关，即局部且微弱的输入触发在特征空间的可分离性有限，部分方法为抑制强后门引入较强正则后更倾向捕获显著特征，从而对微弱后门不够敏感。同时，PTAP在Blend攻击下的检测率较NC和Tabor略有下降，原因在于Blend攻击采用全局混合触发模式，触发信号更易与样本原有特征融合且变化更为微弱，压缩了目标类相对其他类别的扰动强度优势，使异常信号更接近阈值。该现象与图4一致，Blend攻击场景下异常目标标签的异常指数虽超过1.96，但越过阈值的幅度较小且更接近阈值边缘，其扰动强度虽整体低于正常标签，但分布间隔收敛、下移幅度有限。

为进一步解释PTAP的判定依据，图4(a)给出了不同攻击场景下的异常指数及判定阈值，其中Clean对应干净模型的结果，其异常指数取所有标签中最大值，用于刻画在无后门情况下可能出现的最坏波动。可以看到，在多数后门攻击场景中，异常目标标签的异常指数显著高于检测阈值1.96，甚至远高于3.5，该阈值在近似正态假设下可视为对

表 2 不同防御方法在 CIFAR10 和 GTSRB 数据集上的检测率

数据集	攻击方法	NC	FeaTure-RE	Tabor	Pixel	UNICORN	PTAP
CIFAR10	BadNets	**100%**	75%	**100%**	**100%**	90%	**100%**
	Blend	**100%**	100%	**100%**	**100%**	95%	90%
	Bpp	0%	50%	10%	5%	70%	**90%**
	WaNet	45%	60%	35%	40%	**100%**	**100%**
	SSBA	20%	**95%**	30%	35%	90%	**95%**
	DFST	0%	80%	15%	25%	75%	**95%**
	DEFEAT	10%	75%	5%	15%	**90%**	85%
	Trojan	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
	ITI	0%	65%	5%	40%	75%	**100%**
	DFDT	0%	20%	0%	0%	55%	**95%**
	PBADT	5%	40%	0%	35%	80%	**95%**
GTSRB	BadNets	**100%**	65%	**100%**	**100%**	**100%**	**100%**
	Blend	95%	**100%**	95%	**100%**	95%	90%
	Bpp	0%	35%	0%	5%	55%	**85%**
	WaNet	55%	70%	30%	50%	**100%**	90%
	SSBA	15%	85%	15%	55%	65%	**90%**
	DFST	0%	85%	10%	30%	90%	**100%**
	DEFEAT	5%	**85%**	20%	10%	80%	80%
	Trojan	**100%**	**100%**	**95%**	**100%**	**100%**	**100%**
	ITI	0%	85%	0%	40%	70%	**100%**
	DFDT	5%	20%	0%	5%	60%	**100%**
	PBADT	0%	70%	10%	25%	75%	**90%**



(a) 不同攻击场景下的异常指数及判定阈值　　(b) 正常标签与异常目标标签的扰动强度分布

图 4　PTAP 在不同攻击场景下的异常判定情况

应 99.9% 以上的等效置信水平，因此能够以较低误报率筛出显著异常目标。相对地，干净模型下各标签的异常指数整体远低于检测阈值，说明该判定依据在无后门时具有较好的保守性。图 4(b)展示了 CIFAR10 数据集上正常标签与异常目标标签的扰动强度分布差异，其中箱线图由正常标签形成，叉形点表示被标记的异常目标标签。从扰动强度分布上看，异常目标标签对应的扰动强度通常显著小于正常标签，异常点往往落在正常分布的下四分位范围之外，甚至明显低于正常标签的下须位置。因此，

一旦模型存在后门目标类，其后门通路就可以通过更小幅度的参数空间扰动被激活，其他正常类别标签仍需要更大幅度的参数空间扰动才能形成相同程度的标签定向迁移。该现象与后门攻击为目标类时通常预留异常易激活通路的机制一致，这也从侧面解释了 PTAP 利用参数空间扰动强度进行异常识别的有效性。

2) 防御性能

本文在 CIFAR10、GTSRB 和 ImageNet 数据集中进行了实验。表 3 给出了在 BadNets、Blend、Bpp、WaNet、DFST、PBADT 和 ITI 这 7 种攻击下，不同防御方法的防御性能，其中加粗数据表示攻击成功率大于 20% 的方案。在 No Defense 条件下，3 组数据集的 ASR 均接近 100%，后门能够稳定触发。以 CIFAR10 数据集为例，PTAP 在多数攻击下能够在较小 CA 代价下显著降低 ASR，整体表现优于其他防御方法。其平均 ASR 降至 6.99%，CA 相比无防御下降幅度仅小于 5%。相比之下，Fine-pruning 虽能限制去除 Bpp 和 WaNet 的效果，但在 BadNets 与 Blend 上仍有残余 ASR。NC 在 BadNets 上有效，但在 Blend、Bpp、WaNet、PBADT 与 DFST 等攻击上仍维持较高 ASR，呈现明显不稳定性。NAD 与 FeaTure-RE 同样存在对部分攻击有效、对部分攻击失效的现象。I-BAU 在 CIFAR10 数据集上整体较强，但其对大型数据集 ImageNet 鲁棒性不足。需要指出的是，在 ImageNet 数据集上，PTAP 在 WaNet 与 PBADT 攻击下的 ASR 分别为 12.91% 与 14.15%，仍存在残余触发，但相较 No Defense 以及 I-BAU 已显著降低。因此，PTAP 在 3 组数据集上均体现出更好的防御性能，在保持 CA 基本稳定的同时，对多种后门攻击实现了更鲁棒的后门效果压制。

5.3　消融分析

1) 微调数据规模的影响

本文评估了微调可用干净数据规模对 PTAP 性能的影响，并在 BadNets、Bpp 与 Blend 这 3 种攻击下进行验证。实验从 CIFAR10 与 GTSRB 训练集中

表 3　　多种后门攻击下各防御方法的缓解性能对比

数据集	攻击方法	No Defense		Fine-pruning		NC		NAD		FeaTure-RE		I-BAU		PTAP	
		CA ↑	ASR ↓	CA ↑	ASR ↓	CA ↑	ASR ↓	CA ↑	ASR ↓	CA ↑	ASR ↓	CA ↑	ASR ↓	CA ↑	ASR ↓
CIFAR10	BadNets	93.21%	99.32%	92.93%	**87.37%**	89.32%	5.54%	90.32%	2.78%	91.53%	**35.81%**	90.63%	1.22%	92.07%	5.76%
	Blend	92.47%	97.85%	91.52%	**76.01%**	91.53%	**84.62%**	89.49%	4.46%	91.87%	**42.54%**	90.35%	0.52%	90.41%	7.67%
	Bpp	94.11%	99.27%	92.48%	6.16%	91.83%	**93.16%**	90.84%	**83.98%**	91.85%	**90.76%**	89.26%	3.71%	90.94%	4.29%
	WaNet	94.08%	80.53%	92.26%	0.12%	90.37%	**93.36%**	91.34%	9.73%	93.05%	0.37%	91.91%	5.16%	90.35%	11.82%
	DFST	95.50%	99.99%	92.46%	**77.19%**	91.51%	**99.02%**	87.27%	16.56%	91.93%	11.18%	85.25%	**26.46%**	91.39%	14.10%
	PBADT	94.32%	99.06%	92.42%	**41.95%**	91.12%	**83.38%**	91.53%	**53.01%**	91.19%	7.67%	91.54%	2.01%	92.17%	17.54%
	ITI	93.15%	99.80%	92.09%	16.18%	91.48%	**82.31%**	91.26%	8.27%	91.45%	6.14%	91.51%	1.41%	90.01%	3.16%
GTSRB	BadNets	96.82%	99.48%	94.91%	**57.81%**	94.35%	9.12%	95.41%	3.67%	96.09%	**90.07%**	95.35%	0.36%	95.01%	0.51%
	Blend	97.05%	99.21%	95.07%	**69.99%**	93.27%	**91.48%**	89.71%	0.56%	92.07%	**91.90%**	95.65%	0.71%	94.14%	3.16%
	Bpp	97.94%	99.22%	95.85%	2.16%	91.87%	**89.13%**	94.32%	**27.81%**	94.52%	**80.20%**	95.51%	0.19%	94.72%	0.09%
	WaNet	97.43%	99.38%	95.26%	1.82%	93.89%	**96.50%**	95.16%	2.63%	97.01%	11.51%	96.19%	0.15%	93.99%	0.75%
	DFST	98.19%	99.75%	94.15%	**41.82%**	93.99%	**97.12%**	90.15%	**26.10%**	95.08	9.10%	89.28%	**45.31%**	93.20%	6.12%
	PBADT	97.01%	99.16%	96.48%	**21.33%**	96.48%	**81.01%**	95.09%	19.93%	94.61%	8.35%	95.45%	0.61%	94.83%	1.27%
	ITI	97.13%	99.46%	94.44%	13.46%	92.58%	**83.14%**	94.21%	7.34%	96.41%	9.37%	95.52%	0.41%	93.25%	0.38%
ImageNet	BadNets	73.81%	99.26%	62.71%	**52.17%**	72.80%	10.27%	70.36%	17.06%	72.16%	**40.26%**	71.31%	**96.18%**	70.21%	8.12%
	Blend	73.35%	99.12%	63.52%	**44.48%**	72.05%	**80.98%**	69.91%	**65.65%**	72.15%	**92.36%**	62.46%	**97.36%**	70.15%	4.51%
	Bpp	68.82%	99.16%	62.48%	0.15%	70.66%	**82.63%**	61.67%	**21.95%**	62.70%	**94.51%**	60.12%	**90.03%**	71.10%	5.14%
	WaNet	74.02%	99.18%	61.26%	0.42%	72.69%	**87.22%**	70.98%	10.54%	73.19%	8.75%	72.10%	**90.11%**	70.66%	12.91%
	DFST	75.47%	99.30%	62.18%	**39.85%**	72.50%	**91.01%**	63.01%	12.76%	73.03%	19.09%	65.12%	**92.63%**	70.81%	17.16%
	PBADT	73.41%	99.05%	66.48%	**36.37%**	72.25%	**83.51%**	70.53%	**21.62%**	72.87%	18.84%	70.37%	**92.13%**	69.17%	14.15%
	ITI	74.15%	99.41%	68.03%	9.31%	71.18%	**83.14%**	68.56%	12.26%	72.53%	9.70%	69.45%	**91.57%**	71.87%	8.25%

分别抽取0.1%、0.5%、1%、5%和10%的干净样本用于防御后门攻击，结果如图5所示。随着防御数据量增加，PTAP的后门缓解效果持续提升，攻击成功率整体呈单调下降趋势。在不使用干净样本进行微调时，各攻击方法的ASR均接近100%。当仅提供极少量干净样本时，PTAP已能显著降低ASR。当干净样本比例提升至0.1%~0.5%时，多数攻击场景下ASR可降至约10%及以下，且进一步增加防御数据量会带来更稳定的下降。因此，PTAP在有限干净样本条件下仍具备有效的后门缓解能力，适用于防御方对良性数据访问受限的现实设置。

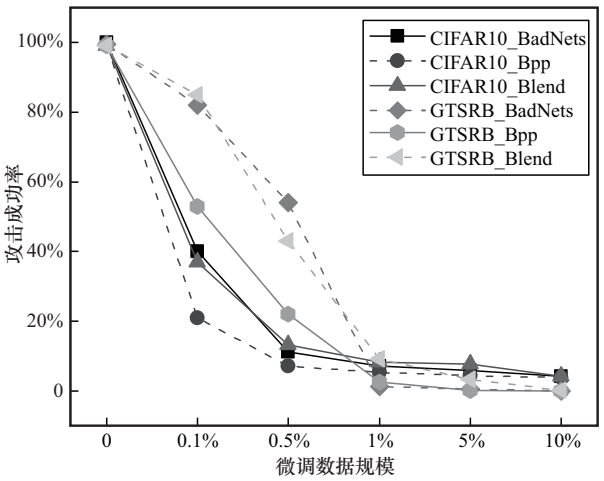


图5 不同微调数据规模下PTAP的攻击成功率变化

2) 权重系数λ的影响

图6展示了在CIFAR10数据集上权重系数λ对PTAP缓解效果的影响。结果表明，λ实际上调节了两种缓解机制的相对强度。当λ较小时，优化过程主要依赖对梯度在后门子空间分量的投影抑制。此时，模型的CA基本保持稳定，但ASR仍有明显残留。这说明仅限制参数更新方向能够阻止后门被进一步强化，却难以快速消除已固化在模型参数中的后门成分。随着λ增大到适中范围，正则约束开始发挥更直接的作用。在持续抑制后门方向更新的同时，权重系数在后门子空间上的投影幅度被逐步压缩，使ASR显著下降，而CA仅出现轻微损失。当λ进一步增大时，ASR的改善趋于饱和，甚至在个别攻击设置下出现小幅回升。与此同时，CA下降更加明显。这是因为过强的正则约束会同时限制与正常判别相关的参数自由度，损害模型的泛化性能。考虑到后门抑制效果与良性性能保持之间的权衡，本文在后续实验中将λ=0.010作为默认设置。

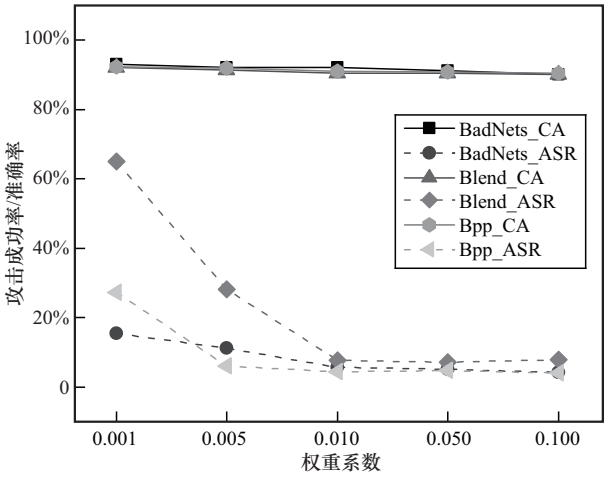


图6 不同权重系数下PTAP的ASR与CA变化

3) 微调学习率的影响

在CIFAR10与ImageNet数据集上，本文针对5.2节微调残留现象，对WaNet、DFST和PBADT进行更强的防御微调，结果如表4所示。由表4可知，增大学习率可显著降低ASR，但会带来CA下降，且该权衡在ImageNet数据集上更为剧烈。PBADT在学习率为0.015时已将ASR降至1.54%，继续增大学习率收益有限但CA进一步下降。对于ImageNet数据集，当学习率从0.010提升至0.015时，ASR大幅下降，同时CA明显下滑。当学习率进一步增至0.020时，DFST的ASR降至0.07%，CA降至65.19%。因此，结合5.2节

表4 不同微调学习率对防御效果的影响

数据集	攻击方法	lr=0.001		lr=0.005		lr=0.010		lr=0.015		lr=0.020	
		CA ↑	ASR ↓	CA ↑	ASR ↓	CA ↑	ASR ↓	CA ↑	ASR ↓	CA ↑	ASR ↓
CIFAR10	WaNet	93.94%	87.44%	93.39%	59.69%	90.35%	11.82%	90.68%	8.78%	90.15%	5.13%
	DFST	94.61%	99.19%	92.50%	87.38%	91.39%	14.10%	91.06%	6.04%	90.83%	4.60%
	PBADT	94.12%	98.66%	92.52%	93.20%	92.17%	17.54%	91.37%	1.54%	90.28%	1.39%
ImageNet	WaNet	73.71%	30.52%	72.56%	22.21%	70.66%	12.91%	66.65%	0.40%	64.78%	0.73%
	DFST	74.29%	53.32%	73.98%	27.47%	70.91%	17.16%	68.31%	4.26%	65.19%	0.07%
	PBADT	73.09%	30.21%	71.48%	19.10%	69.17%	14.15%	65.15%	1.74%	62.28%	0.32%

分析，lr=0.010在多数设置下可提供更均衡的CA与ASR折中。若需进一步抑制ASR，尤其在WaNet与PBADT等攻击方法上，需提高学习率并权衡CA损失。

4) 微调轮数的影响

为分析微调轮数对后门缓解过程的影响，本文在CIFAR10与GTSRB数据集上构造后门模型，并在BadNets与BPP两类典型攻击下评估PTAP微调过程中训练准确率、测试准确率与攻击成功率的动态变化，实验结果如图7与图8所示。ASR通常在前3~5轮快速下降，并在不超过10轮时收敛至较低水平，随后进入平台期，继续训练对ASR的边际收益有限。因此，多数设置下微调轮数可控制在10轮以内，并可据此采用以ASR为核心的早停策略，以降低额外开销并避免无效训练带来的波动。

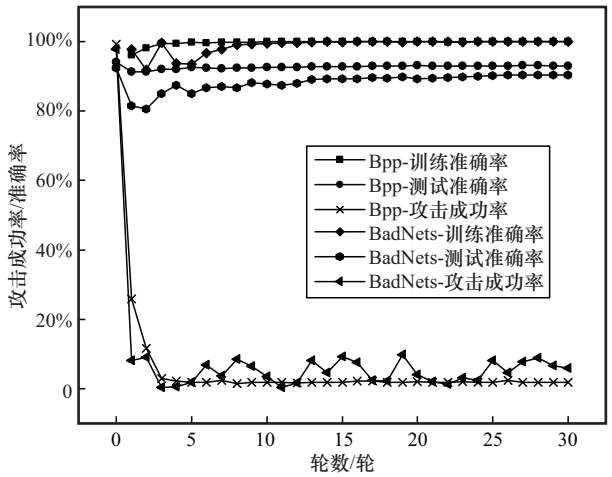


图7　CIFAR10下不同微调轮数的指标变化

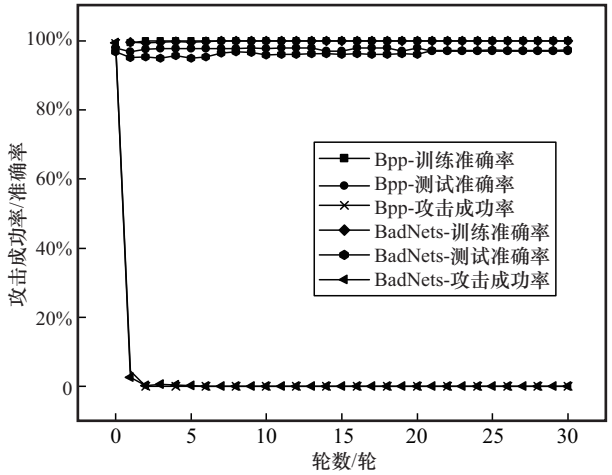


图8　GTSRB下不同微调轮数的指标变化

同时，不同数据集与攻击类型的收敛形态存在差异，GTSRB数据集上微调对两类攻击均呈现快速且稳定的收敛特性。相比之下，CIFAR10_BadNets的ASR虽在早期轮次显著下降，但随后出现多次反弹与振荡，这表明单纯增加轮数并不会带来更稳健的后门抑制效果。CIFAR10_BadNets的波动现象可能与学习率偏大导致更新幅度过大有关，使模型在恢复干净精度时反复调整决策边界并阶段性增强触发响应。因此，微调策略需结合数据集、模型与攻击形态选择合适的学习率与轮数，并配合学习率衰减与早停机制提升稳定性。

5) 约束ρ的影响

为分析成功率阈值ρ对检测稳定性的影响，本文在CIFAR10数据集上分别对BadNets与Bpp两种攻击调节ρ，并统计MAD所标记异常目标标签对应的异常指数，结果如图9所示。随着ρ从0.1提升至0.9，两种攻击的异常指数整体呈上升趋势，且BadNets的增幅更为显著，说明更严格的成功率阈值能够放大后门目标标签在扰动强度集合中的低值离群程度，从而提高异常检测的判别显著性。在低阈值区间，异常指数普遍低于阈值1.96（置信度95%），异常性不够突出，容易削弱目标与普通标签之间的可分性，进而增加漏检风险。当ρ提升至中高区间后，异常指数显著增大并趋于稳定，此时MAD判定更可靠。需要注意的是，ρ过大通常会提高达到成功判据的优化代价，并可能导致部分标签在给定迭代预算内难以达标，从而降低统计量的稳定性并引入波动。因此，ρ的选择需要在异常显著性提升与计算开销及统计稳定性之间取得折中，本文采用中等水平的ρ作为默认配置，以兼顾效率与鲁棒性。

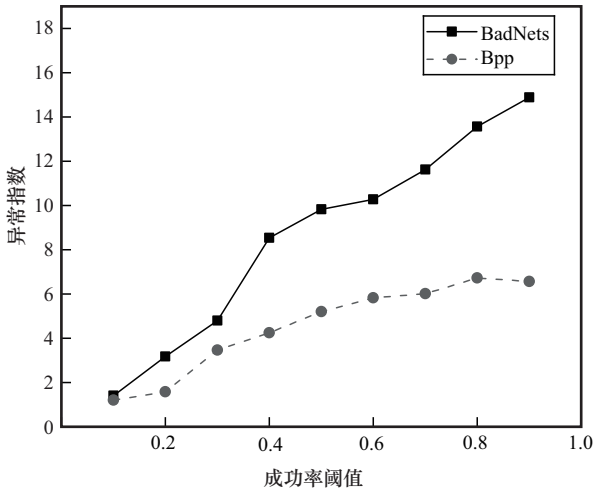


图9　不同成功率阈值下目标标签的MAD异常指数

6) 投毒率的影响

为分析投毒率对单标签后门检测与缓解的影响，本文在CIFAR10与GTSRB数据集上分别构建BadNets与WaNet后门模型，并在不同投毒率下进行异常检测及微调。为进一步压制WaNet残留，微调学习率统一设置为0.015，使用ASR表示微调前的攻击成功率，P-ASR表示微调后的攻击成功率，结果如图10所示。投毒率升高会同步增强后门效应并放大检测信号，检测方面两数据集上的异常值均随投毒率整体上升。表5展示了缓解方面的影响，BadNets在各投毒率下微调后ASR被降至极低水平，WaNet对投毒率更敏感，低投毒率区间呈现检测弱且更易残留的特征，CIFAR10_WaNet在投毒率为0.5%时异常指数仅为2.04，接近阈值边缘。因此，投毒率升高后后门模式更一致，目标离群性更突出，从而提升检测稳健性，并使微调更易将ASR降至较低水平。

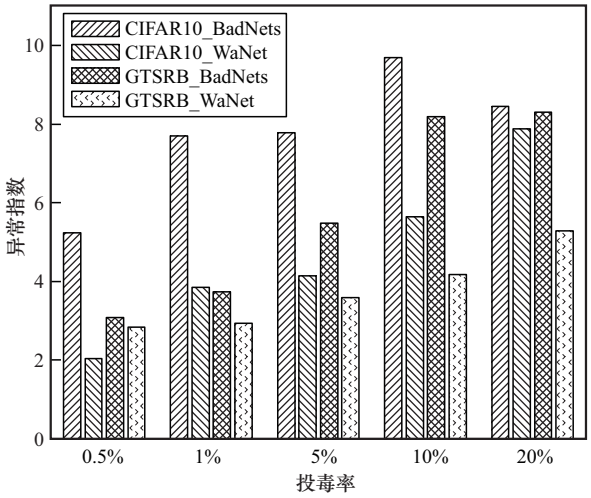


图10 不同投毒率下目标标签的MAD异常指数

5.4 计算开销

本文在CIFAR10与ImageNet数据集上分别统计了不同后门检测方法的计算开销，以评估其效率与可扩展性，结果如图11所示，PTAP在保证检测效果的同时显著降低了计算开销。在CIFAR10数据集的检测实验中，PTAP在BadNets与WaNet攻击下的耗时分别为86.11 s与136 s，均显著低于各类基线方法，其相较UNICORN存在明显数量级差距，同时也稳定快于NC、FeaTure-RE、Tabor与Pixel等代表性检测方法。进一步，本文在更大规模的ImageNet数据集上统计了计算开销。结果显示，PTAP在ImageNet数据集的BadNets与WaNet场景下耗时分别为2 719 s与2 684 s，仍保持在分钟级别。相比之下，NC与Pixel均上升到小时级别。

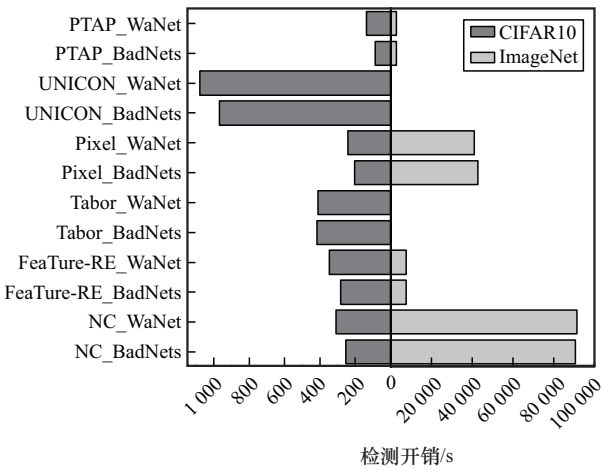


图11 不同检测方法的计算开销

从跨数据集扩展趋势看，输入空间触发反演方法的开销增长更为剧烈。BadNets下NC与Pixel从CIFAR10到ImageNet的耗时分别约增加358倍与208倍，PTAP约增加32倍。因此，输入空间触发反演对数据规模更敏感，PTAP由于在参数空间进行定向对抗扰动，避免样本级搜索与反演开销，对数据规模的依赖相对更弱，在中大规模视觉任务中更具效率优势。

5.5 局限性与讨论

1) 可扩展性分析

① 规模较大的数据集。为验证检测方法在更

表5 不同投毒率对防御效果的影响

数据集	攻击方法	r=0.5%			r=1%			r=5%			r=10%			r=20%		
		CA	ASR	P-ASR	CA	ASR	P-ASR	CA	ASR	P-ASR	CA	ASR	P-ASR	CA	ASR	P-ASR
CIFAR10	BadNets	93.76%	50.06%	2.26%	93.34%	74.73%	4.67%	93.02%	88.74%	2.03%	93.21%	99.32%	1.29%	91.33%	99.05%	1.19%
	WaNet	94.27%	1.12%	0.08%	94.65%	12.64%	1.1%	94.06%	75.50%	7.69%	94.08%	80.53%	8.78%	92.24%	94.71%	4.96%
ImageNet	BadNets	98.73%	68.80%	0.01%	97.89%	93.00%	0.03%	97.62%	98.48%	0.01%	96.82%	99.32%	0.08%	94.90%	99.37%	0.01%
	WaNet	98.67%	60.37%	4.91%	98.25%	59.94%	2.90%	97.62%	92.85%	0.07%	97.43%	99.18%	0.02%	95.61%	98.81%	0.08%

大规模数据集上的可扩展性，本文进一步在 Tiny-ImageNet 与 ImageNet 数据集上开展检测实验。以 ResNet34 网络在两种数据集上分别训练后门模型并进行评估为例，攻击类型覆盖典型静态、动态与自适应触发，选取 BadNets、WaNet 与 PBADT 作为代表，结果如图 12 所示。在 ResNet34 上，无论是 Tiny-ImageNet 还是 ImageNet 数据集，3 类攻击均产生稳定且显著的异常信号，其中，BadNets 与 WaNet 的检测指标始终保持在较高水平。随着数据规模增大，整体检测指标未出现系统性衰减，表明检测方法在更大规模数据集上仍具有良好的鲁棒性与可扩展性。

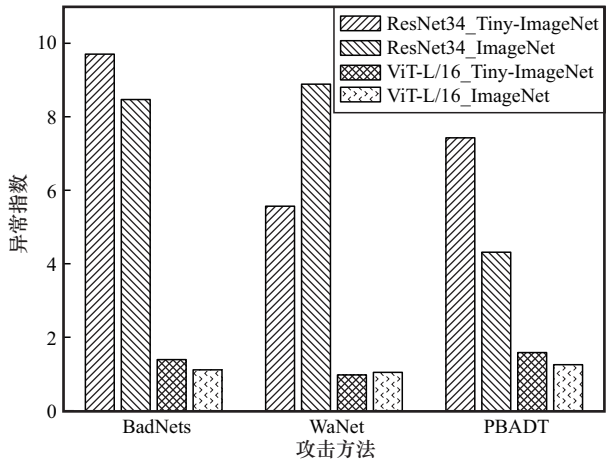


图 12　PTAP 的可扩展性分析

② 规模较大的模型。为评估方法在更大规模视觉模型上的可扩展性，本文选用 ViT-L/16 分别在 Tiny-ImageNet 与 ImageNet 数据集上构造后门模型开展检测实验，如图 12 所示。与中等规模卷积模型上可观察到的显著异常信号不同，ViT-L/16 条件下各攻击方法对应的检测统计量整体降低且趋于收敛，集中于 1~2，未形成稳定的异常离群模式，难以触发高置信度的统计异常。在大型视觉 Transformer 中，后门特征更可能以跨 patch 的分布式形式嵌入表征学习与注意力交互过程，使目标类与非目标类在参数扰动可达性或异常强度上的差距被压缩，从而削弱基于离群性的判别基础。同时，ViT-L/16 参数维度更高，进一步促使各类别统计量趋同并掩盖潜在差异。值得注意的是，相关研究[45]指出，将经典后门检测与触发反演方法直接迁移至 Transformer 或 ViT 往往会出现性能退化或失效的问题，原因在于 ViT 的 patch 处理与注意力机制改变了后门作用形态，使依赖像素局部性假设的检测与防御难以保持有效[46]。因此，ViT-L/16 上的未检出现象更应被视为大型视觉 Transformer 场景中的共性挑战。面向后续改进，可将检测信号从像素局部触发转向 patch 级，围绕注意力响应异常或 patch 级扰动构造更符合 ViT 机理的判别量，或者对参数空间求解进行降维与分层约束，以提升最小扰动估计的可达性与可比性，从而恢复目标类的离群状态。

2) 多后门攻击场景

① 多标签场景。为评估方法在多标签多后门场景下的扩展性，本文在 GTSRB 数据集上构造 BadNets 后门模型。具体地，为 N 个触发器分别生成互异的高饱和度双色棋盘格图案，触发器形状固定为 4×4 方块且位置保持不变。投毒时将样本标签按序映射改写为目标标签序列 1~N，从而形成多目标后门设置，实验结果如图 13 所示。随着触发器数量增加，异常指数持续衰减并呈现平台化趋势。当 N 较小时，目标类仍表现为显著离群。当 N 增至 9（占比 20.93%）及以上时，异常指数迅速降至 1.98，并在 N 为 11~17 时稳定于 1.72~1.85，导致检测难以可靠触发。该现象主要是因为多目标后门集对离群性的稀释，多标签设置将原本集中于单一目标类的后门捷径分散至多个目标类，使多个类别同时呈现不同程度的异常，从而拉低统计分布的中心并改变其离散度，压缩单一目标类相对其余类别的异常间隔，最终削弱基于离群判别的检测有效性。

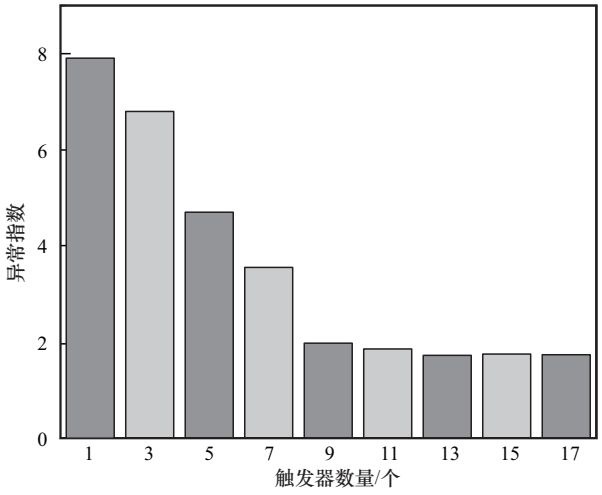


图 13　异常指数随触发器数量的变化

② 单标签多后门场景。为评估方法在单标签多后门场景下的适用性，本文在 GTSRB 数据集上

构造BadNets后门模型，并将目标标签固定为0。具体地，设置触发器数量为9，外观完全一致，均为4×4的纯白方块。触发位置按九宫格覆盖图像空间，包括4个角点、4条边的中点及图像中心，从而在同一目标标签下形成9个位置互异的单标签后门。实验结果如图14所示，初始状态下9个位置触发器的ASR均接近1。随着微调进行，ASR在早期显著下降但呈现位置相关差异，第2轮时多数位置已降至较低水平，但仍存在少数位置残留较高。当微调至第3轮及之后，9个位置的ASR基本降至0，仅有极小残余。因此，在单标签多后门场景下，PTAP能够在有限微调轮次内同步抑制并基本消除多后门效应。

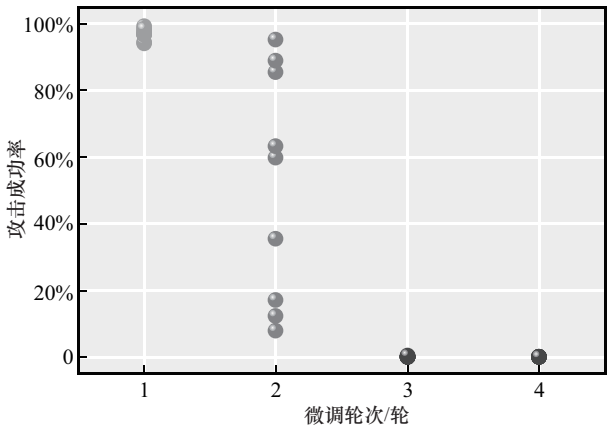


图14 单标签多后门场景下的缓解情况分析

6 伦理声明

本文关注在使用不可信第三方模型或训练资源时可能引入的后门攻击。该类攻击会在用户缺乏对训练流程与数据来源的完整控制时，对模型安全与下游应用造成潜在危害。为降低此类风险，本文方法旨在帮助防御方识别并削弱可疑后门行为，研究目的与技术路线均面向防御，不以提升攻击能力为目标，也未引入新的攻击范式或扩大攻击面。同时需要指出，本文方法的有效性建立在防御能够获得一定数量的本地干净样本用于评估或微调的前提下。在缺乏可信本地数据或风险来自第三方数据污染等其他情形时，后门威胁仍可能存在。对于此类场景，建议优先采用可信来源的模型与数据，并结合更严格的审计、数据筛查与隔离部署策略。因此，本文意在提升模型使用过程中的安全性，而不应被解读为对后门风险已被彻底消除的保证。

7 结束语

为了解决现有后门防御机制对显著且可分离的后门特征的过度依赖，以及触发反转的计算成本的问题，本文提出了一种面向参数空间的对抗扰动方法PTAP。与传统基于输入或特征空间触发反演检测方法不同，PTAP的核心统计量是使每个候选目标类别达到预定义成功率所需的最小参数扰动幅度，以显著降低后门模型检测的计算开销。这些扰动揭示出的异常敏感方向进一步指导了轻量级的微调，从而实现了检测与修复一体化流程。本文在CIFAR10、GTSRB和ImageNet数据集上进行了广泛的比较和消融实验，涵盖了输入空间、特征空间和动态触发设置下的后门攻击。实验表明，PTAP对后门目标的检测置信度超过99%。在缓解措施方面，后门平均成功率降低至6.99%，而主任务准确率的下降幅度保持在5%以下，验证了该方法的有效性和可用性。未来工作将通过参数高效的调整与分层扰动评估相结合，使该框架适应大型序列模型（LLM/Transformer），以支持其向通用化和高可扩展性方向发展。

参考文献：

[1] Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database[C]//Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2009: 248-255.

[2] Parkhi O M, Vedaldi A, Zisserman A. Deep face recognition[C]//Proceedings of the British Machine Vision Conference 2015. Piscataway: IEEE Press, 2015: 1-12.

[3] Yurtsever E, Lambert J, Carballo A, et al. A survey of autonomous driving: common practices and emerging technologies[J]. IEEE Access, 2020, 8: 58443-58469.

[4] Ribeiro M, Grolinger K, Capretz M A M. MLaaS: machine learning as a service[C]//Proceedings of the 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA). Piscataway: IEEE Press, 2015: 896-902.

[5] Jia Y Q, Shelhamer E, Donahue J, et al. Caffe: convolutional architecture for fast feature embedding[C]//Proceedings of the 22nd ACM International Conference on Multimedia. New York: ACM Press, 2014: 675-678.

[6] Koh J Y. Model zoo: discover open source deep learning code and pretrained models[EB]. (2018-06-14)[2026-01-30].

[7] Chen X Y, Liu C, Li B, et al. Targeted backdoor attacks on deep learning systems using data poisoning[PP]. V1. (2017-12-15)[2026-01-30]. arXiv: arXiv.1712.05526.

[8] Voth D, Dane L, Grebe J, et al. Effective backdoor learning on open-set face recognition systems[C]//Proceedings of the 2025 IEEE/CVF Win-

ter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE Press, 2025: 1027-1039.

[9] Pourkeshavarz M, Sabokrou M, Rasouli A. Adversarial backdoor attack by naturalistic data poisoning on trajectory prediction in autonomous driving[C]//Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2024: 14885-14894.

[10] Wang B L, Yao Y S, Shan S, et al. Neural cleanse: identifying and mitigating backdoor attacks in neural networks[C]//Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2019: 707-723.

[11] Guo W B, Wang L, Xu Y, et al. Towards inspecting and eliminating Trojan backdoors in deep neural networks[C]//Proceedings of the 2020 IEEE International Conference on Data Mining (ICDM). Piscataway: IEEE Press, 2020: 162-171.

[12] Xu X Y, Ersoy O, Tajalli B, et al. Universal soldier: using universal adversarial perturbations for detecting backdoor attacks[C]//Proceedings of the 2024 54th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W). Piscataway: IEEE Press, 2024: 66-73.

[13] Wang Z T, Mei K, Ding H L, et al. Rethinking the reverse-engineering of Trojan triggers[C]//Proceedings of the Advances in Neural Information Processing Systems 35. Massachusetts: MIT Press, 2022: 9738-9753.

[14] Xu X, Huang K Z, Li Y M, et al. Towards reliable and efficient backdoor trigger inversion via decoupling benign features[C]//Proceedings of the International Conference on Learning Representations (ICLR). Vancouver: ICLR, 2024: 13784-13809.

[15] Xu X Y, Liu Z R, Koffas S, et al. BAN: detecting backdoors activated by adversarial neuron noise[C]//Proceedings of the Advances in Neural Information Processing Systems 37. Massachusetts: MIT Press, 2024: 114348-114373.

[16] Gu T Y, Dolan-Gavitt B, Garg S. BadNets: identifying vulnerabilities in the machine learning model supply chain[PP]. V2. (2019-03-11)[2026-01-30]. arXiv: arXiv.1708.06733.

[17] Turner A, Tsipras D, Madry A. Label-consistent backdoor attacks[PP]. V2. (2019-12-06)[2026-01-30]. arXiv: arXiv.1912.02771.

[18] Li Y Z, Li Y M, Wu B Y, et al. Invisible backdoor attack with sample-specific triggers[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2021: 16443-16452.

[19] Nguyen T A, Tran T A. Input-aware dynamic backdoor attack[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. New York: ACM Press, 2020: 3454-3464.

[20] Nguyen T A, Tran A T. WaNet: imperceptible warping-based backdoor attack[C]//Proceedings of the International Conference on Learning Representations (ICLR). Vancouver: ICLR, 2021: 6667-6683.

[21] 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 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 15054-15063.

[22] Chen X, Li M H, Sun Y B, et al. Invisible trigger image: a dynamic neural backdoor attack based on hidden feature[J]. Neurocomputing, 2025, 639: 130296.

[23] Cheng S Y, Liu Y Q, Ma S Q, et al. Deep feature space Trojan attack of neural networks by controlled detoxification[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(2): 1148-1156.

[24] Zhao Z D, Chen X J, Xuan Y X, et al. DEFEAT: deep hidden feature backdoor attacks by imperceptible perturbation and latent representation constraints[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 15192-15201.

[25] Li Q Y, Chen W, Xu X T, et al. Precision strike: precise backdoor attack with dynamic trigger[J]. Computers & Security, 2025, 148: 104101.

[26] Chen W M, Xu X W, Wang X D, et al. Dynamic frequency domain trigger backdoor attack with steganography against deep neural networks[J]. Information Sciences, 2025, 718: 122368.

[27] Chen H L, Fu C, Zhao J S, et al. DeepInspect: a black-box Trojan detection and mitigation framework for deep neural networks[C]//Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization, 2019: 4658-4664.

[28] Doan B G, Abbasnejad E, Ranasinghe D C. Februus: input purification defense against Trojan attacks on deep neural network systems[C]//Proceedings of the 36th Annual Computer Security Applications Conference. New York: ACM Press, 2020: 897-912.

[29] Tran B, Li J, Mądry A. Spectral signatures in backdoor attacks[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. New York: ACM Press, 2018: 8011-8021.

[30] Tao G H, Shen G Y, Liu Y Q, et al. Better trigger inversion optimization in backdoor scanning[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 13358-13368.

[31] Hu X L, Lin X, Cogswell M, et al. Trigger hunting with a topological prior for trojan detection[C]//Proceedings of the International Conference on Learning Representations (ICLR). Vancouver: ICLR, 2022: 23417-23434.

[32] Li Y G, Lyu X X, Koren N, et al. Neural attention distillation: erasing backdoor triggers from deep neural networks[C]//Proceedings of the International Conference on Learning Representations (ICLR). Vancouver: ICLR, 2021: 11661-11680.

[33] Zeng Y, Chen S, Park W, et al. Adversarial unlearning of backdoors via implicit hypergradient[C]//Proceedings of the International Conference on Learning Representations (ICLR). Vancouver: ICLR, 2022: 1-27.

[34] Wu D X, Wang Y S. Adversarial neuron pruning purifies backdoored deep models[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems. New York: ACM Press, 2021: 16913-16925.

[35] Li Y G, Lyu X X, Koren N, et al. Anti-backdoor learning: training clean models on poisoned data[C]//Proceedings of the Neural Information Processing Systems (NeurIPS 2021). Massachusetts: MIT Press, 2021: 14900-14912.

[36] Wang Z T, Ding H L, Zhai J, et al. Training with more confidence: mitigating injected and natural backdoors during training[C]//Proceedings of the Advances in Neural Information Processing Systems 35. Massachusetts: MIT Press, 2022: 36396-36410.

[37] Shen G Y, Liu Y Q, Tao G H, et al. Backdoor scanning for deep neural networks through K-arm optimization[C]//Proceedings of the 38th International Conference on Machine Learning (ICML). New York: PMLR, 2021: 9525-9536.

[38] Wang Z T, Mei K, Zhai J, et al. UNICORN: a unified backdoor trigger inversion framework[C]//Proceedings of the International Conference on Learning Representations (ICLR). Vancouver: ICLR, 2023: 26127-26147.
[39] Zhang H T, Wang Y C, Yan S H, et al. Test-time backdoor detection for object detection models[C]//Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2025: 24377-24386.
[40] Zhai S F, Li J J, Liu Y, et al. Efficient input-level backdoor defense on text-to-image synthesis via neuron activation variation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2025: 15182-15193.
[41] Wu B Y, Chen H R, Zhang M D, et al. BackdoorBench: a comprehensive benchmark of backdoor learning[C]//Proceedings of the Advances in Neural Information Processing Systems 35. Massachusetts: MIT Press, 2022: 10546-10559.
[42] Hampel F R. The influence curve and its role in robust estimation[J]. Journal of the American Statistical Association, 1974, 69(346): 383-393.
[43] Liu Y Q, Ma S Q, Aafer Y, et al. Trojaning attack on neural networks[C]//Proceedings 2018 Network and Distributed System Security Symposium. Internet Society, 2018: 18-21.
[44] Liu K, Dolan-Gavitt B, Garg S. Fine-pruning: defending against backdooring attacks on deep neural networks[C]//Research in Attacks, Intrusions, and Defenses. Berlin: Springer, 2018: 273-294.
[45] Raj R, Roy B, Das A, et al. “We must protect the transformers”: understanding efficacy of backdoor attack mitigation on transformer models[C]//Security, Privacy, and Applied Cryptography Engineering. Berlin: Springer, 2024: 242-260.
[46] Yuan Z H, Zhou P, Zou K, et al. You are catching my attention: are vision transformers bad learners under backdoor attacks? [C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2023: 24605-24615.

[作者简介]

田有亮（1982– ），男，贵州盘州人，博士，贵州大学教授，主要研究方向为博弈论、密码学与安全协议。

金昆龙（2000– ），男，贵州盘州人，贵州大学硕士生，主要研究方向为隐私保护、联邦学习、后门攻击。

石璐嘉（2000– ），女，重庆人，贵州大学硕士生，主要研究方向为人工智能、模型水印等。

王帅（2000– ），男，贵州贵阳人，贵州大学博士生，主要研究方向为隐私保护、联邦学习、安全聚合等。

左建烁（2002– ），男，河南新乡人，贵州大学硕士生，主要研究方向为人工智能、图像取证等。

向阿新（1996– ），男，贵州遵义人，博士，贵州大学副教授，主要研究方向为区块链、密码学、密钥管理等。