计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(9): 2386−2400, 2026

联邦学习自适应梯度累积后门攻击

张桂芬[1,3,4] 苏秦春[2,3,4] 李鸿梓[2,3,4] 葛丽娜[1,3,4]

[1](广西民族大学人工智能学院 南宁 530006)

[2](广西民族大学物理与电子信息学院 南宁 530006)

[3](智能计算网络与大数据信息安全重点实验室(广西民族大学) 南宁 530006)

[4](广西混杂计算与集成电路设计分析重点实验室(广西民族大学) 南宁 530006)

(zhangguifen@gxmzu.edu.cn)

Federated Learning Adaptive Gradient Accumulation Backdoor Attack

Zhang Guifen[1,3,4], Su Qinchun[2,3,4], Li Hongzi[2,3,4], and Ge Lina[1,3,4]

[1](*School of Artificial Intelligence*, *Guangxi Minzu University*, *Nanning* 530006)

[2](*School of Physics and Electronic Information*, *Guangxi Minzu University*, *Nanning* 530006)

[3](*Key Laboratory of Intelligent Computing Network and Big Data Information Security* (*Guangxi Minzu University*), *Nanning* 530006)

[4](*Guangxi Key Laboratory of Hybrid Computation and IC Design Analysis* (*Guangxi Minzu University*), *Nanning* 530006)

Abstract Adaptive gradient accumulation backdoor attack (AGABA) framework is designed for federated learning (FL) systems, integrating a parameterized adaptive subblock trigger (AST) with dynamic transparency and a multistage gradient accumulation (MGA) mechanism. This framework effectively addresses the persistent trade-off dilemma between stealthiness and persistence that plagues traditional backdoor attacks in the FL environment. AST decomposes the complete trigger into multiple independent components via dynamic transparency control and distributed subblock superposition technology, allowing malicious clients to collaboratively construct a global trigger pattern while maintaining a high level of attack stealthiness. MGA mechanism adopts a three-phase attack strategy including initial accumulation, gradient accumulation and attack execution, and combines it with a parameter importance-aware mechanism. It realizes the latent hiding and targeted activation of malicious updates in model aggregation through cross-round progressive gradient accumulation. Moreover, AGABA adopts momentum-accelerated gradient divergence propagation and adaptive memory factor adjustment, ensuring that all malicious gradient updates fall within the legitimate distribution interval of the FL system and thus evading statistical anomaly-based detection mechanisms effectively. Experiments show that with 20% malicious clients participating in FL training, AGABA can maintain a favorable backdoor attack success rate even under the protection of various mainstream defense mechanisms, and its overall performance outperforms single backdoor attack methods.

Key words federated learning; distributed; centralized; backdoor attacks; gradient accumulation; adaptive attacks

摘 要 联邦学习系统的自适应梯度累积后门攻击(adaptive gradient accumulation backdoor attack, AGABA)框架结合了动态透明度的参数化自适应子块触发器(adaptive subblock trigger, AST)和多阶段梯度累积(multistage gradient accumulation, MGA)机制,有效解决了传统后门攻击在联邦环境下面临的隐蔽性

收稿日期: 2025-08-27; **修回日期**: 2026-05-07

基金项目: 国家自然科学基金项目(61862007); 广西自然科学基金项目(2024GXNSFAA010111)

This work was supported by the National Natural Science Foundation of China (61862007) and the Guangxi Natural Science Foundation (2024GXNSFAA010111).

通信作者: 葛丽娜(19920028@gxmzu.edu.cn)

与持久性平衡难题。AST通过动态透明度控制和分布式子块叠加技术，将完整触发器分解为多个独立组件，使恶意客户端能够在保持高度隐蔽性的同时协同构建全局触发模式。MGA采用3阶段攻击策略（初始累积、梯度累积和攻击实施）结合参数重要性感知机制，通过跨轮次的渐进式梯度累积实现恶意更新在模型聚合中的潜伏与激活。该框架利用动量加速的梯度差异传播和自适应记忆因子调整，确保攻击梯度始终位于合法分布区间内，有效规避基于统计异常的检测机制。实验表明，在20%恶意客户端参与的场景下，AGABA能够在多种主流防御机制保护下保持较好的后门攻击成功率，优于现有单一攻击方法。

关键词 联邦学习；分布式；集中式；后门攻击；梯度累积；自适应攻击

中图法分类号 TP391

DOI: 10.7544/issn1000-1239.202550637 **CSTR:** 32373.14.issn1000-1239.202550637

联邦学习（federated learning, FL）作为一种新型的分布式机器学习方法，其核心优势在于保护数据隐私情况下实现跨设备的知识共享。相较于传统集中式需要上传原始数据的模式，FL仅要求参与方共享本地训练的模型参数，从根本上规避了数据泄露风险。这一特性使得FL在医疗健康、物联网及金融科技等领域展现出独特价值[1-2]。随着实际应用场景的深入拓展，研究者发现FL在数据异构性、通信效率与安全层面暴露出瓶颈。数据异构导致模型适用性受限[3]，复杂网络拓扑结构加剧节点间通信开销[4]，分布式协作机制本身更衍生出新型安全威胁，恶意客户端可通过模型参数投毒操纵全局模型决策，梯度交换过程潜藏成员隐私推理风险，而后门攻击因具备强隐蔽性与目标定向性尤为突出，攻击者通过精心设计的触发模式污染本地训练过程，使得全局模型对特定输入产生预设的异常响应。

早期后门攻击聚焦于数据层面的显式触发器植入，例如，Gu等人[5]系统揭示了深度神经网络的后门脆弱性，通过在训练集中注入带有特定触发器的中毒样本（如像素块标记），使模型在测试阶段对含触发器的输入产生预定错误分类。然而，此类显式攻击易被基于输入筛查的防御机制检测。为提升隐蔽性，Saha等人[6]提出隐式触发器嵌入方案，利用图像隐写技术规避视觉审查，但这些方法均建立在攻击者完全掌控训练数据的假设之上，难以适应FL场景的分布式特性与数据隔离约束。FL的多方参与特性大幅增加了攻击面，Sun等人[7]系统论证了FL系统在数据中毒攻击下的脆弱性，但集中式触发策略存在单点失效风险，其可靠性和鲁棒性仍有待提高。为了克服攻击的局限性，Xie等人[8]充分利用FL特性，提出了DBA分布式后门攻击方法，将全局触发器分解为多个局部组件，实现了攻击负载的分散化，但其静态触发器设计仍存在被检测风险。而Liu等人[9]通过排列组合策略进一步增强了DBA的攻击效果，但缺乏对防御机制的自适应能力。

后门攻击在FL环境下面临三大技术挑战：1）隐蔽性。恶意客户端比例有限，异常梯度更新容易被鲁棒聚合机制识别并过滤。2）持久性。单轮次攻击强度不足以在全局模型中留下稳定后门，而持续高强度攻击又会增加被检测风险。3）协同性。分布式环境下缺乏攻击者间的直接协调机制，难以实现一致性后门目标的植入。现有研究受限于静态触发器设计与单轮攻击思路，在面对先进联邦防御机制时表现脆弱。

针对以上问题，本文提出自适应梯度累积后门攻击（adaptive gradient accumulation backdoor attack, AGABA）框架，具有以下技术创新和实验贡献：

1）自适应子块触发器（adaptive subblock trigger, AST）。设计了动态透明度调整策略的触发器生成机制，能够根据时间维度（训练轮次进展）、防御维度（检测状态感知）和攻击效果维度（成功率反馈）自适应调整触发器特征，实现早期隐蔽、后期增效的动态攻击模式。

2）多阶段梯度累积（multistage gradient accumulation, MGA）。提出初始化—累积—攻击3阶段转换策略，设计阶段转换条件和动态记忆因子，在不同攻击阶段平衡历史梯度与当前梯度的影响权重，确保攻击方向的一致性与累积效应的最大化。

3）参数重要性感知。建立基于扰动分析的参数重要性评估机制，实现重要参数重点攻击、次要参数轻微扰动的攻击策略，在提升攻击效率的同时降低异常特征，增强对防御机制的规避能力。

4）对抗防御评估。系统评估了AGABA对主流FL防御机制的突破能力，揭示了现有防御在面对分

布式累积攻击时存在的不足。

1 相关工作

1.1 后门攻击

在FL后门攻击研究中，数据层面的后门触发器可以分为集中式和分布式2种形式，相较于集中式触发器，分布式后门攻击模式通过分散化的触发机制能有效提高攻击的隐蔽性，文献[8]提出的DBA分布式触发器概念，为后续研究奠定了基础。文献[9]则在此基础上提出FCBA，进一步增强了触发器的效果。然而，这些触发器在面对动态防御机制时的适应能力方面仍存在明显局限。

在进一步后门攻击中，参数层面的后门攻击也不断发展。Bagdasaryan等人[10]提出了对抗性参数扰动技术，通过精心优化后门参数更新的分布特征，使其在统计特性上与良性更新高度匹配，有效规避了基于异常检测的防御机制。Zhang等人[11]提出Neurotoxin方法，通过在训练轮次中注入微小但持续的后门扰动，借助局部模型更新的累积效应，实现了攻击的高持久性和隐蔽性。Fang等人[12]则从另一角度提出了符号翻转攻击框架，该方法通过直接修改本地模型权重的符号来注入后门，避开传统的触发数据训练路径。但这些方法在攻击持久性和跨轮次一致性方面表现不足，尤其是在分布式环境中难以协调多个恶意客户端形成统一攻击方向。

因此，本文在数据触发器层面，提出AST方法，通过动态透明度调整和随机子块激活机制，赋予各恶意客户端时变特性，实现全局触发器隐式协同聚合。在参数扰动层面，提出了MGA方法，采用3阶段渐进式攻击策略，将攻击负载分散至多轮次、多客户端。这2种策略形成了一个完整的攻击体系，AST负责生成高隐蔽性的视觉触发器，而MGA则保证攻击梯度在参数空间的有效累积与分散，共同克服了单一攻击方法的局限性，提高了FL后门攻击的隐蔽性、持久性和环境适应能力。

1.2 后门防御

当前，FL中的后门防御研究主要沿着模型细化、鲁棒性聚合和认证鲁棒性3条技术路线展开。

1)模型细化方法。通过对训练过程和参数空间的精细调控，消除潜在后门影响。这类方法不直接排除恶意客户端，而是试图修复可能被污染的模型。FedDF[13]通过知识蒸馏技术，融合多个本地模型的核心知识，降低后门特征在蒸馏过程中的传递效率。FedMV[14]，Fine_Tuning[15]则分别从剪枝方法和参数微调角度，减轻后门特征对模型决策边界的扭曲。

2)鲁棒聚合方法。专注于优化FL的核心环节——模型聚合过程，通过识别并过滤恶意更新来保护全局模型。Rflbat[16]，Robustlr[17]引入了基于更新统计特征的异常检测机制，对可疑客户端进行降权处理。Krum[18]，Median[19]则采用几何距离或统计中位数原则，选择最具代表性的客户端更新进行聚合。FLTrust[20]通过建立信任评分系统，为每个客户端贡献分配可信度权重。Flame[21]，Bulyan[22]则引入更复杂的多轮筛选机制，通过递进式过滤剔除异常更新。

3)认证鲁棒性方法。从理论层面为FL系统提供安全保障，通过严格的数学证明确保即使在后门攻击存在的情况下，模型性能和安全性仍能满足预设标准。CRFL[23]提出了一个2级认证框架，结合服务器端验证和客户端证明，实现了可量化的FL鲁棒性保证。

2 自适应梯度累积后门攻击

2.1 威胁模型

在c个客户端的FL中，有m个客户端($m<c$)被攻击者控制，攻击者的目标是在保持全局模型对正常样本准确率的前提下，使模型在遇到特定触发器时将输入样本分类为预设目标类别。攻击者能力限制如下：

1)控制范围。攻击者操控的m个恶意客户端可以自由修改其本地数据集、训练过程和上传的模型更新。攻击者在受控客户端中引入特制的触发样本，构造混合数据集D_{mal}，生成携带后门特征的有害梯度。

2)系统状态不可见。无法获取FL训练过程中全局模型权重、全局梯度和学习率，攻击者无法干扰这些系统级配置，只能在受控客户端中遵循全局一致的超参数设置进行本地训练。

3)客户端隔离保障。攻击者无法越界访问任何客户端，每个良性客户端的训练环境对攻击者而言均为封闭的黑盒，其训练过程独立进行，仅基于自身数据集和服务器分发的全局模型参数。攻击者的影响范围严格限定在被控制的恶意客户端内部操作，通过聚合机制间接影响全局模型的性能表现。

2.2 攻击框架概述

如图1所示的FL训练示意图，AGABA攻击框架中包含AST和MGA这2个核心组件，前者负责生成数据层面的分布式触发器，后者控制参数空间的梯度渐进累积。

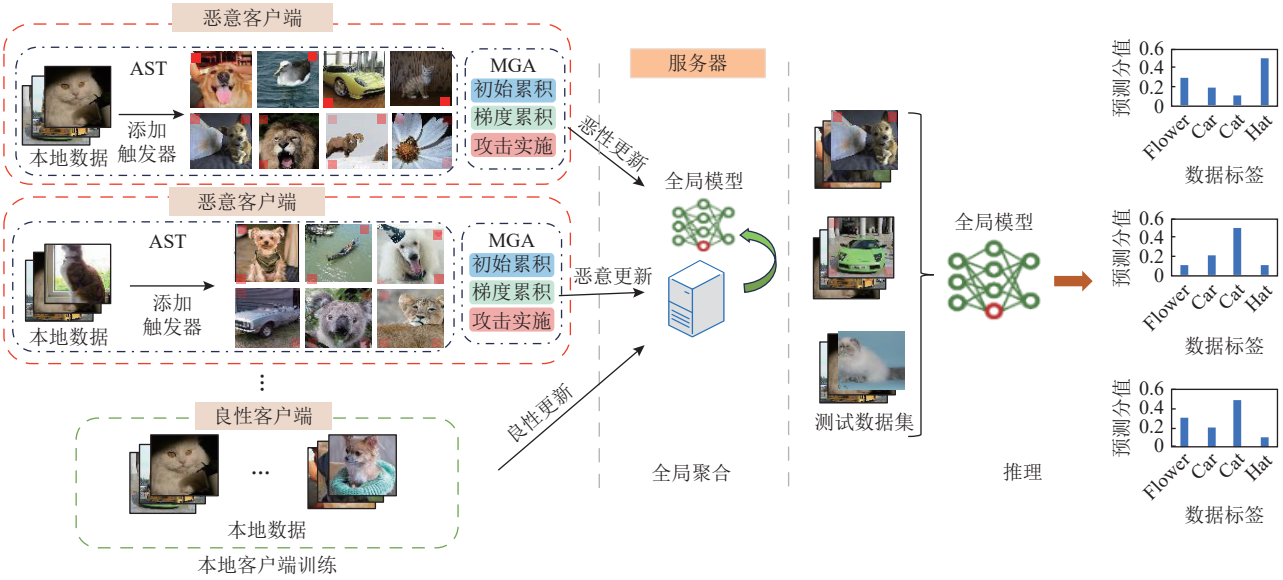


Fig. 1 Illustration of federated learning training

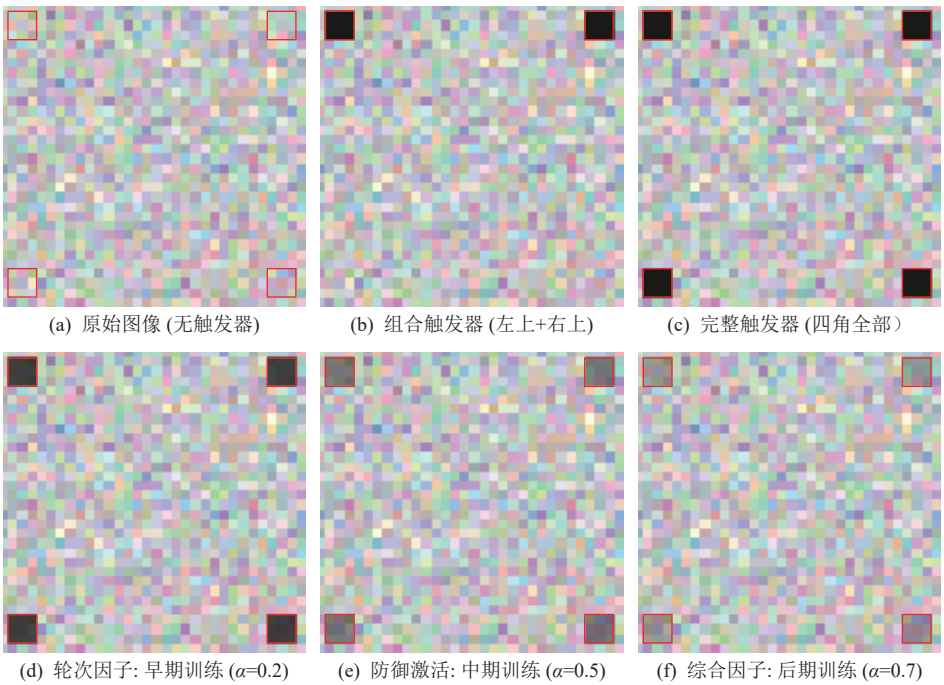
图 1 联邦学习训练示意图

恶意客户端通过 AGABA 框架协同发起后门攻击，每个攻击阶段的实施策略如下：1)初始阶段，各恶意客户端建立全局梯度状态共享机制，计算并存储参数重要性映射；2)累积阶段，多轮次渐进式累积攻击梯度同时动态调整透明度和记忆因子；3)攻击阶段，当累积梯度达到触发阈值时，将攻击载荷注入全局模型，实现后门植入。整个过程中，恶意客户端通过跨轮次、跨客户端的协同工作，确保攻击方向的一致性和攻击效果的持久性。

2.3 AST

如图 2 所示的是 AST 部分示意图，AST 位置使用红色边框进行标明，其中 α 为透明度因子。AST 将触发器模式分解为多个独立的子块，每个子块可独立激活并应用于输入样本的不同空间位置。这种设计通过动态透明度控制的叠加机制，在保证攻击有效性的同时实现了高度隐蔽的分布式触发效果。

AST 的触发器子块集合定义为式(1)，由多个局部触发模式组成：



(a) 原始图像 (无触发器) (b) 组合触发器 (左上+右上) (c) 完整触发器 (四角全部)

(d) 轮次因子: 早期训练 (α=0.2) (e) 防御激活: 中期训练 (α=0.5) (f) 综合因子: 后期训练 (α=0.7)

Fig. 2 Illustration of AST section

图 2 AST 部分示意图

$$B=\{b_1,b_2,\cdots,b_n\}\,, \tag{1}$$

其中b_i表示第i个触发子块，n表示最大子块数量。在实际攻击过程中，每个恶意客户端从子块集合B中随机选择若干子块进行组合，形成具有随机性的触发器变体，降低被模式识别算法检测的风险。

AST的核心在于多因子自适应透明度控制机制，其透明度因子α控制原始图像与触发器的混合比例，如式(2)所示。

$$\alpha=\alpha_{\text{base}}\cdot f_{\text{epoch}}\cdot f_{\text{defense}}\cdot f_{\text{success}}\,, \tag{2}$$

其中α_{base}为基础透明度取值范围，f_{epoch}，f_{defense}，f_{success}分别为轮次因子、防御检测因子和成功率因子。

轮次因子设计用于随着训练的进行逐渐降低透明度，使攻击更加隐蔽，具体计算如式(3)所示。

$$f_{\text{epoch}}=1-\frac{t}{T}\cdot\gamma_t\,, \tag{3}$$

其中t表示当前训练轮次，T表示总训练轮次，γ_t表示轮次调整权重。确保攻击在训练初期具有较强的攻击效果，在训练后期具有较高的隐蔽性。防御检测因子根据是否检测到防御机制的存在动态调整透明度。当检测到防御机制时，增加透明度以增强攻击效果，如式(4)(5)所示。

$$f_{\text{defense}}=\begin{cases}1.5, & \text{if } detected=\text{True}\,,\\ 1.0, & \text{其他}\,,\end{cases} \tag{4}$$

$$\text{防御检测值}detected=\frac{rejected_update}{total_update}>\tau\,, \tag{5}$$

其中$rejected_update$表示被拒绝更新次数，$total_update$表示总更新次数，τ为检测阈值。成功率因子根据当前攻击成功率(*success_rate*)动态调整透明度，平衡攻击效果和隐蔽性，当攻击成功率低于预期时，增加透明度以增强攻击效果；当攻击成功率已经很高时，降低透明度以增强隐蔽性，如式(6)所示。

$$f_{\text{success}}=\begin{cases}1.3, & \text{if } success_rate<0.4\,,\\ 1.0, & \text{if } 0.4\leqslant success_rate\leqslant 0.8\,,\\ 0.8, & \text{if } success_rate>0.8\,。\end{cases} \tag{6}$$

将输入图像的过程采用透明度叠加机制，颜色叠加如式(7)所示。

$$I_{\text{new}}=\alpha\cdot I_{\text{original}}+(1-\alpha)\cdot C\,, \tag{7}$$

其中I_{new}表示在指定区域叠加触发器后的图像，I_{original}表示原始图像，C表示触发器的颜色。具体算法流程如算法1所示。

算法 1. AST生成。

输入：原始图像I_{original}，数据集 *dataset*，测试标志 *test_mode*，叠加触发器块数 *max_blocks*；

输出：添加触发器的图像I_{new}。

① 初始化*dataset* ←选择触发器颜色 C、位置策略 P；

/*计算动态透明度因子*/

② $\alpha\leftarrow\alpha_{\text{base}}\cdot f_{\text{epoch}}\cdot f_{\text{defense}}\cdot f_{\text{success}}$；

③ if *test_mode* = True then

/*所有触发器子块*/

④ *selected_blocks* ← B；

⑤ else

/*随机选择子块*/

⑥ *selected_blocks* ← *Random*(B)；

⑦ end if

/*初始化输出图像*/

⑧ $I_{\text{new}}\leftarrow I_{\text{original}}$；

⑨ for each b_i in *selected_blocks* do

⑩ $I_{\text{new}}\leftarrow\alpha\cdot I_{\text{original}}+(1-\alpha)\cdot C$；

⑪ end for

⑫ return I_{new}。

2.4 MGA后门攻击

MGA后门攻击策略是将攻击载荷以渐进方式嵌入模型中，在多个训练轮次中缓慢积累攻击信号，避免一次性攻击带来的异常梯度，从而规避FL中常见异常检测机制。

如图3所示，MGA包含初始累积(initial)阶段、梯度累积(accumulation)阶段和攻击(attack)实施阶段。

相较于传统后门攻击直接对模型施加大幅度的梯度扰动，MGA引入3个关键模块：参数重要性加权，针对神经网络中对后门触发器敏感的关键参数进行定位和加权处理；自适应梯度累积，基于攻击阶段和成功率动态调整梯度累积策略；记忆因子控制，通过记忆因子平衡历史累积梯度与当前梯度的比例，在保证攻击有效性的同时提高隐蔽性。

参数重要性加权模块，衡量损失函数在参数扰动前后的变化，可以近似估计每个参数对模型输出的影响程度。给定损失函数L，其在参数p处的泰勒展开为

$$L(p+\delta)=(p)+\frac{\partial L}{\partial p}\delta+\frac{1}{2}\frac{\partial^2 L}{\partial p^2}\delta^2+o(\delta^3)\,。 \tag{8}$$

当扰动δ足够小时，可近似得到参数重要性，如式(9)所示。

$$I(p)\approx\left|\frac{\partial L}{\partial p}\right|\cdot|\delta|\,, \tag{9}$$

其中$I(p)$表示参数p的重要性，说明参数p对损失函数$L(p)$的影响，梯度越大，参数越重要。因此MGA首先通过扰动分析计算模型各参数对后门攻击的重

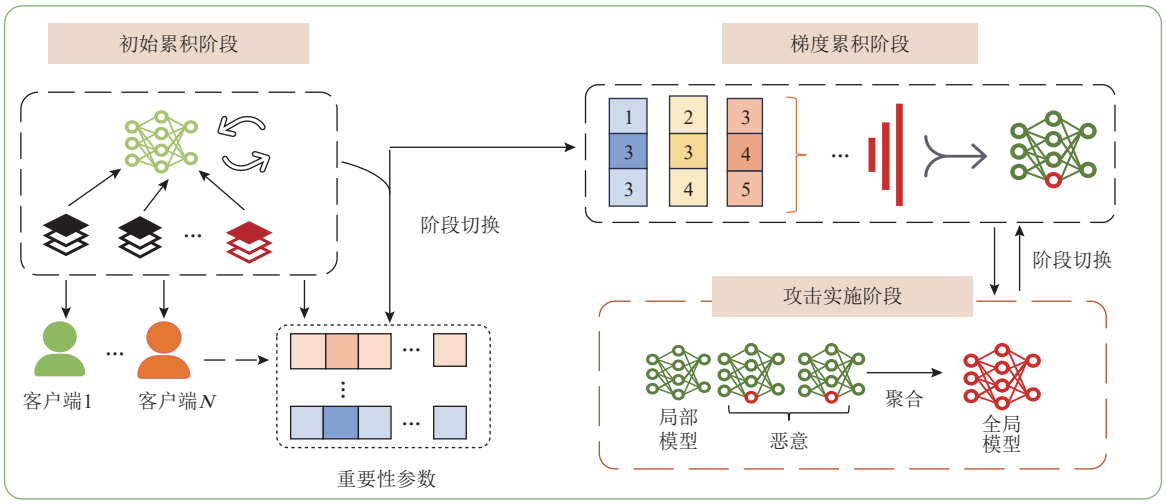


Fig. 3 MGA attack strategy

图 3 MGA 攻击策略

要性，如式(10)所示。

$$I(p)=|L(M(x,p+\delta),y)-L(M(x,p),y)|, \tag{10}$$

其中M表示模型，x表示输入样本，y表示目标标签，δ表示微小高斯扰动。计算扰动前后损失函数的差值作为该参数的重要性指标，在获得所有参数的重要性评分后，对其归一化处理，保持值域在 [0, 1] 区间内，如式(11)所示。

$$I'(p)=\frac{I(p)}{\max\{I(p_i)|p_i\in\mathcal{P}\}}, \tag{11}$$

其中$I'(p)$表示归一化后的重要性指标，用于放大对后面触发器敏感参数的贡献，$\mathcal{P}$表示模型参数集合。

自适应梯度累积机制在每轮本地训练中，攻击者根据当前梯度和历史累积梯度按式(12)进行更新累积：

$$G_{\text{accum}}^{(t)}=\beta_t\cdot G_{\text{accum}}^{(t-1)}+(1-\beta_t)\cdot I'(p)\cdot G_{\text{curr}}^{(t)}, \tag{12}$$

其中$G_{\text{accum}}^{(t)}$表示第 t 轮的累积梯度，β_t表示第 t 轮的记忆因子，控制历史信息与当前更新的权重，$G_{\text{accum}}^{(t-1)}$表示历史累积梯度，$G_{\text{curr}}^{(t)}$表示第 t 轮计算得到的梯度。为保证收敛性，MGA 设定梯度上界$\|G_{\text{curr}}^{(t)}\|\leqslant K$，由于重要性指标满足$I'(p)\leqslant 1$，则有式(13)不等式成立。

$$\left\|G_{\text{accum}}^{(t)}\right\|\leqslant\beta_t\left\|G_{\text{accum}}^{(t-1)}\right\|+(1-\beta_t)K。 \tag{13}$$

当$t\to\infty$时，累积梯度有界，如式(14)所示。

$$\lim_{t\to\infty}\left\|G_{\text{accum}}^{(t)}\right\|\leqslant K。 \tag{14}$$

在持续累积的过程中，梯度幅度依然可控，不会因累积造成异常突变。

证明. 设$a_t=\left\|G_{\text{accum}}^{(t)}\right\|$，$\beta_t=\beta$为常数，则有递归关系：

$$a_t\leqslant\beta a_{t-1}+(1-\beta)K。 \tag{15}$$

解这个不等式，得到式(16)：

$$\begin{aligned}a_t&\leqslant\beta^t a_0+(1-\beta)K\sum_{i=0}^{t-1}\beta^i,\\&\leqslant\beta^t a_0+(1-\beta)K\cdot\frac{1-\beta^t}{1-\beta},\\&\leqslant\beta^t a_0+K(1-\beta^t)。\end{aligned} \tag{16}$$

当$t\to\infty$时，由于$0<\beta<1$，有$\beta^t\to 0$，因此

$$\lim_{t\to\infty}a_t\leqslant K, \tag{17}$$

a_t有界。 证毕.

记忆因子调节策略，在不同攻击阶段设定不同的记忆因子β_t，以动态平衡历史与当前梯度，如式(18)所示。

$$\beta_t=\begin{cases}0.9, & \text{if } phase=\text{initial},\\0.7, & \text{if } phase=\text{accumulation},\\0.6, & \text{if } phase=\text{attack},\end{cases} \tag{18}$$

其中较高的记忆因子值会赋予历史梯度更大的权重，保持攻击方向的一致性；较低的记忆因子值则更重视当前梯度，增强攻击能力。*phase* 为阶段切换，MGA 的 3 个阶段攻击流程与切换详述如下：

1)初始累积阶段。建立稳定的攻击方向，避免梯度振荡。此阶段攻击方向尚不稳定，采用较高的记忆因子(β=0.9)，需要更多地依赖历史梯度平滑，初始阶段通常持续到累积步数达到 5 或检测到初步攻击成功为止，攻击初步生效时，进入累积阶段。累积步数是指特定攻击阶段内执行梯度累积的次数，而非全局训练轮次。

2)梯度累积阶段。平衡历史与当前梯度，渐进调整攻击方向，采用中等记忆因子($\beta=0.7$)，动态调整攻击策略，逐步增强累积梯度强度。当累积步数超过 3 或平均攻击成功率超过 60% 时，表明攻击已达到理想效果，即系统进入最终的攻击实施阶段。

3)攻击实施阶段。此阶段的主要目标是最大化攻击成功率,同时保持足够的隐蔽性。降低记忆因子($\beta=0.6$),更加注重当前梯度的即时影响,释放累积攻击载荷,将后门特征嵌入全局模型。为了保持攻击的持续有效性,系统持续监控攻击成功率,当成功率下降到30%以下时,会回退到梯度累积阶段重新积累攻击梯度,表明当前累积的攻击梯度已经失效或被防御机制削弱,需要重新进入累积阶段以恢复攻击能力,这种动态调整机制确保了攻击的持续性和适应性。

在攻击实施阶段,累积梯度(G_{accum})将被用于参数更新,如式(19)所示。

$$M_{t+1}=M_t-\lambda\cdot G_{\text{accum}}^{(t)}, \tag{19}$$

其中M_t表示第t轮的模型,λ表示攻击强度调节因子,用于控制最终注入强度。MGA具体算法流程如算法2所示。

算法 2. MGA攻击策略。

输入:模型M,本地样本数据x,目标标签y,最大轮次T;

输出:攻击后的模型M'。

/*初始化*/

① $phase\leftarrow$"initial",$\beta\leftarrow 0.9$,$G_{\text{accum}}\leftarrow 0$;

② for $t=1$ to T do

③ for 每个恶意客户端 do

④ $x'\leftarrow$调用AST生成(x);

/*计算参数重要性*/

⑤ if $phase=$"initial" then

⑥ for 每个参数p do

⑦ $I(p)\leftarrow$计算重要性参数;

⑧ end for

⑨ $I'(p)\leftarrow\dfrac{I(p)}{\max\{I(p_i)|p_i\in\mathcal{P}\}}$;

⑩ end if

/*更新累积梯度*/

⑪ $G_{\text{accum}}^{(t)}\leftarrow\beta_t\cdot G_{\text{accum}}^{(t-1)}+(1-\beta_t)\cdot I'(p)\cdot G_{\text{curr}}^{(t)}$;

⑫ if $phase=$"attack" then /*应用累积梯度*/

⑬ $M_{t+1}\leftarrow M_t-\lambda\cdot G_{\text{accum}}^{(t)}$;

⑭ end if

⑮ end for

⑯ end for

⑰ return M'。

3 实验分析

本文实验环境是Python模拟了FL场景,具体配置为:Intel® Xeon® Platinum 8352V CPU(2.10 GHz),RTX 5090(32 GB),Linux操作系统,Python 3.8.20,PyTorch 1.11.0+cu113。

客户端的总数量为100,每轮训练随机选择20个客户端参与全局聚合,恶意客户端数占参与聚合的客户端数20%,中毒率20%。数据集的分布考虑了独立同分布(independent and identically distributed, IID)和非独立同分布(non-independent and identically distributed, Non-IID)两种情况,在Non-IID场景中,使用Dirichlet分布进行采样,其参数默认设置为0.5。

实验选取了CIFAR-10和Fashion-MNIST数据集进行测试,模型架构分别采用了ResNet-18和2Conv+2fc。CIFAR-10包含50 000张训练图像和10 000张测试图像,均为32×32像素,涵盖10个类别;Fashion-MNIST数据集包含60 000张训练图像和10 000张测试图像,图像尺寸为28×28像素,涵盖10个类别。

AGABA框架在3种模型细化防御[13-15]、7种鲁棒聚合防御[16-22]和1种认证防御[23]下评估,并与DBA[8],FCBA[9],Neurotoxin[11]后门攻击进行对比分析。

3.1 评估指标

后门攻击的性能通过以下2个关键指标进行衡量:后门攻击成功率(attack success rate, *ASR*)和主任务准确率(accuracy, *ACC*)。*ASR*是指带毒样本被中毒模型分类为目标攻击标签的比例,用于衡量后门攻击的有效性;*ACC*是指模型对干净样本的分类能力。定义如式(20)和式(21)所示。

$$ASR=\frac{N_{\text{attack}}}{N_{\text{trigger}}}\times 100\%, \tag{20}$$

其中N_{attack}是被模型成功分类为目标攻击标签的带触发器样本数量,N_{trigger}是总的带触发器样本数量。

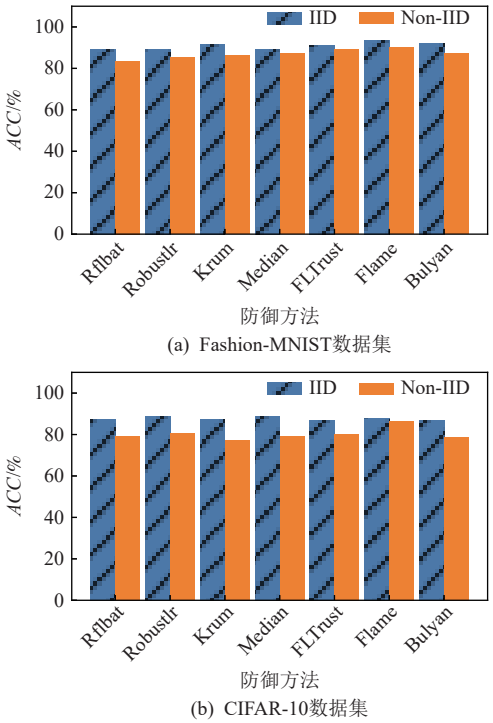
$$ACC=\frac{N_{\text{correct}}}{N_{\text{clean}}}\times 100\%, \tag{21}$$

其中N_{correct}是模型在干净测试集上分类正确的样本数量,N_{clean}是干净测试集中的总样本数量。

3.2 无攻击下的防御效果

为了分析后门攻击对模型性能的影响,本研究采用FL中广泛使用的FedAvg方法作为标准基线,其并入细化防御中,与后续的多种防御机制进行对比评估。

1)鲁棒聚合方法的*ACC*如图4所示。在IID和Non-IID数据分布下,Flame在Fashion-MNIST中表现最佳,得益于其基于模型聚类与动态裁剪的鲁棒聚合机制能有效过滤恶意更新。Krum和Median性能略有下降,尤其是在CIFAR-10的Non-IID场景下,*ACC*



(a) Fashion-MNIST数据集

(b) CIFAR-10数据集

Fig. 4 Impact of robust aggregation defense without attacks on *ACC*

图 4 无攻击的鲁棒聚合防御对 *ACC* 的影响

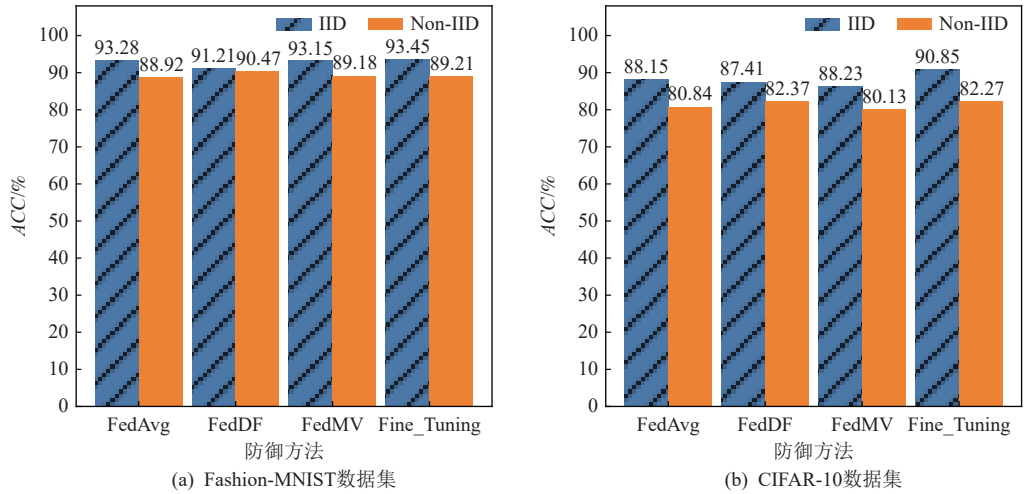
分别为 77.32% 和 79.19%，这可能是在处理数据异质性时，依赖剪枝策略可能削弱模型的鲁棒性。

2）无攻击的细化防御效果如图 5。细化防御方法在 IID 数据分布下 *ACC* 表现稳定，在 Fashion-MNIST 中普遍高于 91.21%，在 CIFAR-10 中为 86.23%~90.85%。Non-IID 场景下，由于数据异质性，*ACC* 性能有所下降，在 Fashion-MNIST 中为 89.18%~90.47%，在 CIFAR-10 中为 80.13%~82.37%。其中，FedDF 在 Non-IID 数据中的 *ACC* 表现最佳，在 Fashion-MNIST 中达到 90.47%，在 CIFAR-10 中达到 82.37%，表明其动态知识蒸馏机制对数据分布偏移具有较强适应性。

3）认证鲁棒性防御的 *ACC* 如图 6 所示。在 IID 数据分布下 *ACC* 同样表现良好，在 Fashion-MNIST 中为 93.12%，在 CIFAR-10 中为 89.56%。在 Non-IID 场景下，*ACC* 性能略有下降，在 Fashion-MNIST 中为 89.21%，在 CIFAR-10 中为 82.52%，但仍然保持较高水平，这表明 CRFL 的认证机制在不同数据分布下均能有效维护模型性能。

3.3 防御下的攻击 *ACC* 和 *ASR*

1）表 1 与表 2 展示了 4 种攻击方法在鲁棒聚合防御下的 *ACC* 和 *ASR* 表现。总体来看，AGABA 依然在多数防御策略下保持高 *ASR* 与良好 *ACC*，表现出极强的对抗鲁棒性。CIFAR-10 的 Non-IID 设置中，AGABA 在 Rflbat, Robustlr 下的 *ASR* 分别达到 69.57% 和 71.21%，优于所对比的攻击，Median 防御虽能一定程度压制，但 *ASR* 仍达 81.69%，而 Krum 对所有攻击的抑制作用最强，AGABA 的 *ASR* 也降至 60.43%，但其 *ACC* 高达 83.37%，说明攻击仍可隐蔽执行。DBA 采用静态触发器并在多个客户端协同嵌入，虽然在模型细化中有效，但在鲁棒聚合下效果大幅下降，以 Krum 为例，其核心思想是选择距离最近的一组更新，剔除偏离多数派的异常梯度，而 DBA 中的后门梯度因模式统一、方向集中，很容易被判定为异常并排除。Neurotoxin 通过微调重要参数方向，使模型训练偏向错误目标，其优势在于低可检测性，但在 Median, Robustlr, FLtrust 等防御中，攻击效果明显下降。Median 聚合直接对每个参数取中值，极端值难以生效，Robustlr 根据全局统计调整聚合权重，而 FLtrust 按梯度方向



(a) Fashion-MNIST数据集

(b) CIFAR-10数据集

Fig. 5 Impact of refined defense without attacks on *ACC*

图 5 无攻击的细化防御对 *ACC* 的影响

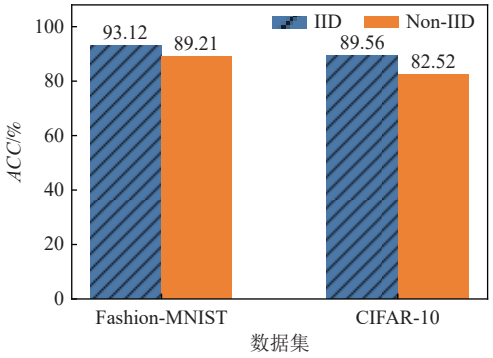


Fig. 6 Impact of authentication robust defense without attacks on *ACC*

图 6 无攻击的认证鲁棒防御对 *ACC* 的影响

可信度加权进一步压制了 Neurotoxin 这种仅修改部分权重方向的攻击。这些传统攻击手段本质上缺乏对防御机制的适应能力，多数是固定式或单维度的扰动策略，面对设计精巧、基于距离、中值或信任机制的鲁棒聚合策略时，往往会被识别并抑制。

2)如表 3 所示为模型细化防御的结果，Fashion-MNIST 数据集在整体来看，4 种攻击均能达到 70%~90% 的精度水平，这表明攻击在模型细化防御下，都能在一定程度上保持模型的预测准确性。

ASR 的表现则存在差异，DBA 攻击在多种主流联邦学习防御机制下，*ASR* 受到明显抑制，往往在 20% 以下波动。DBA 固定触发器受限于防御扰动，FCBA 排列组合策略虽提高 DBA 攻击效果，但也易被细化防御中和，Neurotoxin 虽能影响权重方向，但其不考虑全局训练阶段与防御策略的交互，因此效果不稳定。模型细化防御类方法本质上通过额外训练清洗梯度污染或恢复正常表示，能抵御大多数静态或单一策略攻击，但面对 AGABA 这种动态、自适应、协同式攻击时效果有限。

3)在表 4 的认证鲁棒防御实验中，AGABA 攻击在 2 种数据集中均达到较高的 *ASR*，例如在 Fashion-MNIST 的 IID 下的 *ASR* 高达 92.89%，且 *ACC* 维持在 89.36%，显示其在保留模型性能的同时具备强大的后门植入能力。相比之下，DBA，Neurotoxin 攻击的有效性均被削弱，Neurotoxin 在 CIFAR-10 的 Non-IID 设置中的 *ASR* 仅有 58.66%。CRFL 其核心思想是通过对每轮模型更新进行可认证性分析，例如约束其在聚合过程中的扰动幅度、偏移方向与信赖区间等，以识别和削弱可能的异常更新，因此对强烈突变、明显偏离的攻击具备较强检测与压制能力。Neurotoxin 虽然采用权重方向干预的方式试图提升隐蔽性，但由于缺乏跨轮次自适应调节和与认证边界的对齐机制，仍然可能在某些局部参数上引发偏移突变而被识别和降权。而 AGABA 则通过 AST 生成动态透明

Table 1 Attack *ACC*/*ASR* Under Robust Aggregation Defense

表 1 鲁棒聚合防御下的攻击 *ACC*/*ASR* %

攻击方法	数据集	数据分布	Rflbat		Robustlr		Krum		Median	
			ACC	*ASR*	*ACC*	*ASR*	*ACC*	*ASR*	*ACC*	*ASR*
DBA	Fashion-MNIST	IID	77.31	64.05	84.76	87.87	75.51	60.89	79.38	65.51
		Non-IID	64.35	50.74	81.54	70.49	65.45	54.83	76.3	63.75
	CIFAR-10	IID	79.16	71.66	61.74	65.18	76.19	58.11	79.61	79.55
		Non-IID	63.88	66.39	59.43	62.1	73.93	57.4	76.56	73.38
FCBA	Fashion-MNIST	IID	78.13	66.75	85.05	88.54	76.01	61.78	80.04	64.2
		Non-IID	69.88	58.47	81.8	86.71	69.1	58.38	80.55	64.31
	CIFAR-10	IID	80.16	74.27	68.54	74.18	78.1	60.19	80.72	80.66
		Non-IID	65.88	67.32	60.43	68.1	76.54	59.54	81.34	75.9
Neurotoxin	Fashion-MNIST	IID	77.28	63.02	67.57	54.76	76.26	62.55	76.25	62.5
		Non-IID	68.32	58.86	73.41	59.61	69.29	54.58	79.87	54.47
	CIFAR-10	IID	65.89	69.39	68.62	55.92	48.72	39.66	44.7	33.74
		Non-IID	63.65	67.27	51.7	41.65	51.34	45.9	41.37	43.99
AGABA (本文)	Fashion-MNIST	IID	**81.51**	**67.9**	**86.36**	**89.22**	**77.87**	**62.96**	**83.27**	**68.29**
		Non-IID	**76.09**	**66**	**82.3**	**89.01**	**72.75**	**60.99**	**81.87**	**77.7**
	CIFAR-10	IID	**82.16**	**81.62**	**73.23**	**77.63**	**80.12**	**63.26**	**84.7**	**83.82**
		Non-IID	**73.48**	**69.57**	**67.42**	**71.21**	**77.12**	**60.43**	**83.37**	**81.69**

注：黑体数值为最优值。

Table 2 Attack *ACC*/*ASR* Under Robust Aggregation Defense

表 2 鲁棒聚合防御下的攻击 *ACC*/*ASR* %

攻击方法	数据集	数据分布	FLtrust		Flame		Bulyan	
			ACC	*ASR*	*ACC*	*ASR*	*ACC*	*ASR*
DBA	Fashion-MNIST	IID	90.37	90.82	76.52	61.76	76.56	62.41
		Non-IID	81.23	88.31	68.79	54.79	70.02	57.96
	CIFAR-10	IID	81.95	73.55	61.61	62.16	81.73	77.33
		Non-IID	75.17	75.12	61.25	60	71.94	66.67
FCBA	Fashion-MNIST	IID	90.11	93.96	76.6	50.56	76.09	61.38
		Non-IID	83.31	93.72	74.61	47.13	64.12	57.01
	CIFAR-10	IID	82.82	75.78	78.48	68.26	83.34	71.03
		Non-IID	70.98	77.39	76.51	62.13	72.96	69.38
Neurotoxin	Fashion-MNIST	IID	76.54	62.24	84.88	**63.53**	**78.68**	63.42
		Non-IID	65.98	59.24	77.21	57.04	70.46	57.36
	CIFAR-10	IID	75.79	61.43	58.91	61.22	67.97	50.29
		Non-IID	49.96	31.16	54.94	52.54	68.75	48.85
AGABA（本文）	Fashion-MNIST	IID	**90.91**	**95.57**	**86.5**	62.74	77.47	**68.95**
		Non-IID	**87.77**	**95.6**	**82.75**	**64.47**	**72.59**	**64.31**
	CIFAR-10	IID	**86.77**	**78.94**	**85.69**	**76.44**	**85.81**	**80.11**
		Non-IID	**76.1**	**79.09**	**79.96**	**74.4**	**76.78**	**72.73**

注：黑体数值为最优值。

度、随机位置与分块组合的可调触发器，在图像层面维持微小扰动，避免分布偏移。

4 消融实验

前述实验在理想化可控环境中验证了 AGABA 框架的理论有效性。然而，真实的 FL 系统通常对攻击者呈现为黑盒或灰盒条件，其信息获取严重受限。为评估 AGABA 在实际约束下的鲁棒性与失效边界，如表 5 所示，攻击场景分为白盒、半黑盒和全黑盒，本节将从资源约束和信息约束维度展开分析。消融环境数据集为 Fashion-MNIST，数据分布为 Non-IID，防御方法为 FedMV，中毒样本比例为 20%，恶意客户端比例为 20%。

图 7(a)中 3 种场景下的 *ACC* 均在 20 轮后稳定在 85%~87%，图 7(b)白盒场景中，攻击者可获取大部分状态信息，使 AST 模块能精确计算触发器透明度。MGA 模块根据全局 *ASR* 动态切换相位，这种反馈使攻击快速收敛，25 轮达到 85% 以上，40 轮后逐渐收敛。

在半黑盒场景中，由于本地信息与全局状态存在滞后与偏差，导致其攻击调整呈现波动，*ASR* 在 20~40 轮震荡。在黑盒场景缺乏反馈的情况下，固定策略攻击无法自适应模型的训练动态，其效率低，但

Table 3 Attacks *ACC*/*ASR* Under Model Refinement Defense

表 3 模型细化防御下的攻击 *ACC*/*ASR* %

攻击方法	数据集	数据分布	FedAvg		FedDF		FedMV		Fine_Tuning	
			ACC	*ASR*	*ACC*	*ASR*	*ACC*	*ASR*	*ACC*	*ASR*
DBA	Fashion-MNIST	IID	90.22	92.79	76.82	62.81	86.64	89.73	90.84	78.06
		Non-IID	85.12	90.42	76.19	56.98	82.45	85.64	83.76	79.26
	CIFAR-10	IID	82.4	79.38	79.35	77.16	79.23	80.88	89.37	78.72
		Non-IID	72.63	78.63	71.31	70.93	72.73	75.12	80.7	79.52
FCBA	Fashion-MNIST	IID	90.17	92.99	74.1	62.04	90.41	90.05	92.53	77.85
		Non-IID	87.78	90.69	79.13	62.21	84.1	85.41	84.99	77.38
	CIFAR-10	IID	86.14	85.78	82.87	61.83	82.16	85.35	89.11	78.64
		Non-IID	75.46	77.05	76.57	68.78	74.03	78.91	81.08	78.24
Neurotoxin	Fashion-MNIST	IID	88.71	90.84	60.52	59.19	80.69	52.46	**93.42**	77.4
		Non-IID	84.69	88.42	60.42	52.64	60.98	60.37	86.55	77.86
	CIFAR-10	IID	76.77	72.81	80.42	65.83	75.99	61.38	89.41	79.39
		Non-IID	74.85	77.1	75.11	50.78	62.97	56.56	80.06	79.21
AGABA（本文）	Fashion-MNIST	IID	**91.52**	**93.53**	**76.06**	**70.07**	**90.29**	**91.37**	92.87	**81.44**
		Non-IID	**88.82**	**91.73**	**85.84**	**73.56**	**86.82**	**89.77**	**87.56**	**79.29**
	CIFAR-10	IID	**87.44**	**84.46**	**83.19**	**80.16**	**84.16**	**87.35**	**90.12**	**83.72**
		Non-IID	**76.3**	**83.26**	**77.65**	**76.93**	**76.53**	**81.91**	**82.08**	**80.52**

注：黑体数值为最优值。

Table 4 Attack *ACC*/*ASR* Under Authenticated Robust Defense

表 4 认证鲁棒防御下的攻击 *ACC*/*ASR* %

攻击方法	数据集	数据分布	*ACC*	*ASR*
DBA	Fashion-MNIST	IID	87.89	90.93
		Non-IID	80.01	87.46
	CIFAR-10	IID	84.87	77.28
		Non-IID	79.52	76.29
FCBA	Fashion-MNIST	IID	**90.22**	91.85
		Non-IID	85.93	89.34
	CIFAR-10	IID	87.5	81.78
		Non-IID	**82.12**	75.51
Neurotoxin	Fashion-MNIST	IID	65.34	52.15
		Non-IID	70.98	56.6
	CIFAR-10	IID	76.31	62.2
		Non-IID	63.88	58.66
AGABA（本文）	Fashion-MNIST	IID	89.36	**92.89**
		Non-IID	**88.87**	**91.79**
	CIFAR-10	IID	**88.43**	**85.16**
		Non-IID	81.63	**77.97**

注：黑体数值为最优值。

策略本身因固定而避免了因错误反馈导致的波动，表现稳定缓慢直到 78 轮后 *ASR* 才达到 82% 左右。

4.1 资源约束

1）恶意客户端比例消融，中毒样本比例为 20%。如表 6 所示，恶意客户端比例临界阈值约 10%，低于此阈值时 *ASR* 急剧下降，AGABA 无法在极低恶意比例下生效。在 2% 比例下，后门信号在聚合时被良性更新，完全稀释白盒 AST 因检测到持续低 *ASR* 而不断增大α，反而削弱了本已微弱的触发器信号，而全黑盒的固定策略则避免了这种自我削弱，体现了一定的鲁棒性。在 20%~30% 比例，3 种场景的后门信号强度足以对抗聚合稀释，*ASR* 略有提升但边际收益递减，已接近饱和。

2）中毒样本比例消融，恶意客户端比例为 20%。如表 7 所示，在原始攻击信号极度微弱的场景下，依赖于该信号进行反馈的自适应策略（白盒、半黑盒）容易失效甚至产生负优化，而简单的固定策略反而可能保持一定效果，但后门梯度信号弱，即使通过 MGA 放大也难以对抗主任务梯度。8%~10% 中等投毒比例使后门信号与主任务信号达到平衡，*ASR* 快速提升至 75%~89%。15%~25% 高投毒比例进入饱和区，*ASR* 接近 90% 但 *ACC* 开始下降，因为过多投毒样本可能导致模型过拟合后门任务而忽略主任务。

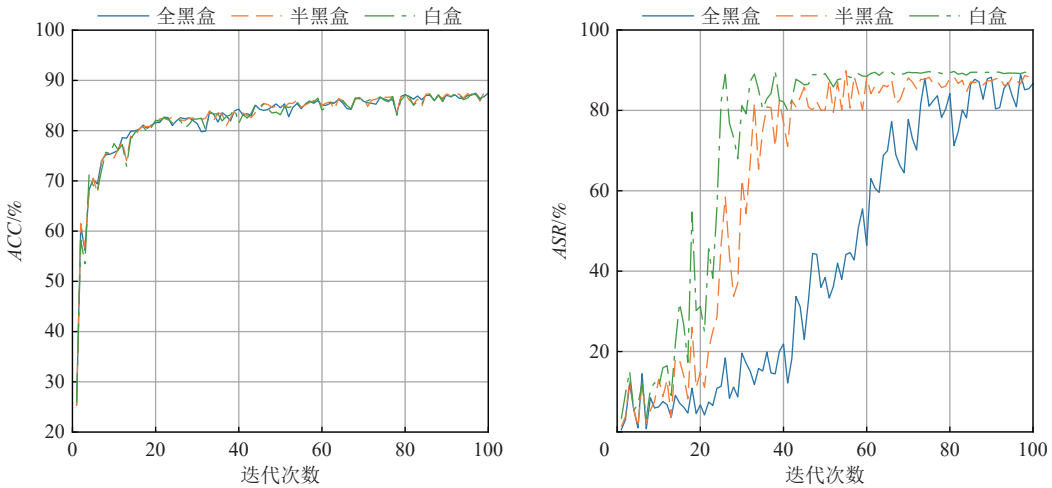
4.2 信息约束

1）AST 组件消融实验。本组实验固定 MGA 策

Table 5 Attack Scenario Settings

表 5 攻击场景设置

场景	AST 透明度策略	MGA 阶段切换策略
白盒	3 因子动态调整 $\alpha=\alpha_{\text{base}}\cdot f_{\text{epoch}}\cdot f_{\text{defense}}\cdot f_{\text{success}}$	基于 *ASR* 切换初始进入累积（*ASR* > 30%），累积进入攻击（*ASR* > 60%）
半黑盒	本地轮次递减 $\alpha=\alpha_{\text{base}}\cdot(1-\min(t/50,0.5))$	本地切换损失 *loss* < 0.8时进入累积，*loss* < 0.3时进入攻击
全黑盒	固定透明度 $\alpha=0.85$	固定轮次切换，轮次< 5为初始，5 ~ 15轮为累积，15 轮之后为攻击



(a) 不同场景下的*ACC*变化 (b) 不同场景下的*ASR*变化

Fig. 7 Comparison curves of *ACC* and *ASR* under three attack scenarios

图 7 3 种攻击场景下的 *ACC* 和 *ASR* 对比曲线

Table 6 Comparison of Malicious Client Proportion Under Three Attack Scenarios

表 6 3 种攻击场景下恶意客户端比例对比 %

场景配置	2%		10%		15%		20%		30%	
	ACC	*ASR*	*ACC*	*ASR*	*ACC*	*ASR*	*ACC*	*ASR*	*ACC*	*ASR*
白盒	88.24	5.74	86.48	81.65	87.23	89.25	87.44	89.42	86.57	89.77
半黑盒	87.88	3.61	86.07	72.33	87.60	85.41	87.83	87.12	87.12	89.46
全黑盒	88.09	6.25	86.62	86.12	87.62	86.68	87.47	88.97	86.55	89.75

Table 7 Comparison of Malicious Sample Proportions Under Three Attack Scenarios

表 7 3 种攻击场景下恶意样本比例对比 %

场景配置	2%		8%		10%		15%		25%	
	ACC	*ASR*	*ACC*	*ASR*	*ACC*	*ASR*	*ACC*	*ASR*	*ACC*	*ASR*
白盒	87.81	34.76	87.43	79.06	85.84	89.02	86.40	89.37	84.65	89.75
半黑盒	87.85	13.91	87.77	75.18	86.51	84.71	87.00	87.22	86.19	89.13
全黑盒	86.96	39.60	87.61	78.33	85.69	88.42	86.81	89.33	85.88	89.76

略为 *ASR* 阶段切换，逐步剥离 AST 透明度计算公式中的动态调整因子。如图 8(a)所示，AST-full 透明度策略使用完整 3 因子，取得了最高且最稳定的 *ASR* 曲线。依次去除成功率因子 AST-nosuccess、防御因子 AST-nodefense 后，*ASR* 的收敛速度与稳定峰值均出现下降，表明每个因子都对攻击优化有正向贡献。当仅保留轮次因子 AST-onlyepoch 时，*ASR* 进一步降低；而使用完全固定的透明度 AST-fixed 时，*ASR* 表现最差，收敛缓慢且波动较大。这一对比表明 AST 模块的核心并非源于某个单一因子，动态调整能力基于多源信息反馈。

2)MGA 组件消融实验。本组实验固定 AST 为完整的 3 因子策略，对比了 MGA 模块不同的阶段切换策略。如图 8(b)所示，MGA-full 基于 *ASR* 阶段切换(阈值 30% 和阈值 60%)，*ASR* 快速上升并稳定在高位。MGA-localloss 本地损失切换(阈值 0.8 和阈值 0.3)效果次之，MGA-fixed 固定轮次切换 (阈值 5 和阈值 15)，MGA- periodic 周期切换(10 轮 1 个周期，前 6 轮累积，后 4 轮攻击)表现稍弱，固定或周期性的策略由于与模型内部状态脱节，无法确保时机上的优化，导致累积的梯度在低效时机释放或攻击力度不足。

3)AGABA 变体消融实验。本组实验对比了完整框架与移除关键组件的变体。如图 8(c)所示，完整 AGABA 框架 full-AGABA 的 *ASR* 曲线优于其他变体；仅使用 AST 的 only-AST 方案而无 MGA，其 *ASR* 增长缓慢，原因在于缺乏梯度累积放大机制，单轮微弱的后门信号在联邦平均聚合中极易被稀释。仅使用 MGA 而固定触发器 only-MGA 的方案，其 *ASR* 曲线起点低且增长不稳定，这是因为 MGA 虽然能放大梯度信号，但固定的触发器无法生成高质量、自适应的后门梯度。传统后门攻击基线 baseline-DBA 效果稍差。综合 3 组消融实验，AST 与 MGA 组件之间协同效应，共同构成了 AGABA 框架实现高效隐蔽的后门攻击。

4)后门遗忘性实验。如图 9 所示，攻击在 50, 100, 150, 200 轮停止，总轮次为 500 时，所有曲线在攻击停止后均呈现下降趋势，持久性的差异本质在于攻击者能否利用反馈信息，将后门梯度引导至与主任务梯度相关性较低的参数子空间。白盒攻击使累积的梯度 G_{accum} 沿着模型损失函数中关于后门任务的某个特异性方向演化，实现了这种隐式的梯度方向优化。当攻击停止，正常训练重启时，主任务产生的梯度 G_{main} 主要作用于对主任务敏感的参数方向，如果后门梯度 G_{accum} 的方向与 G_{main} 的方向具有正交性，即

$$\cos\theta = \frac{\langle G_{\text{accum}}, G_{\text{main}} \rangle}{\| G_{\text{accum}} \| \cdot \| G_{\text{main}} \|} \approx 0 , \tag{22}$$

那么后续的主任务更新对已植入的后门参数扰动影响甚微，后门得以留存。白盒攻击通过反馈机制，近似地优化了这一过程。相反，黑盒攻击由于固定策

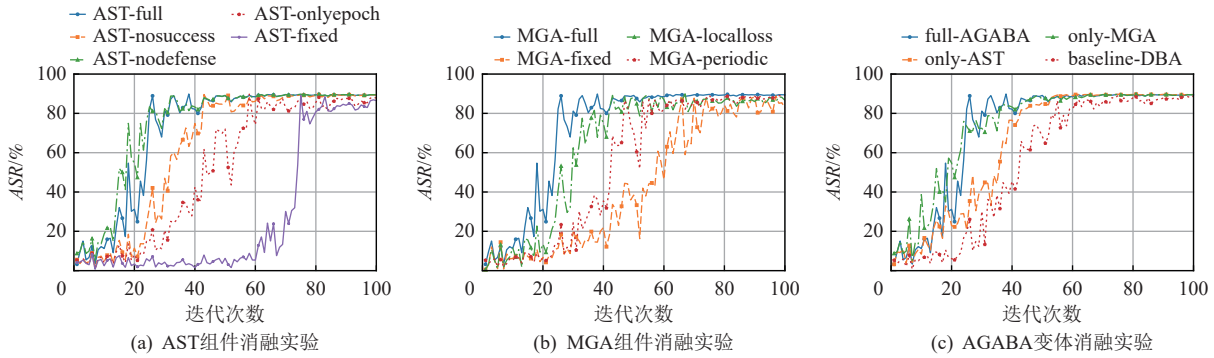
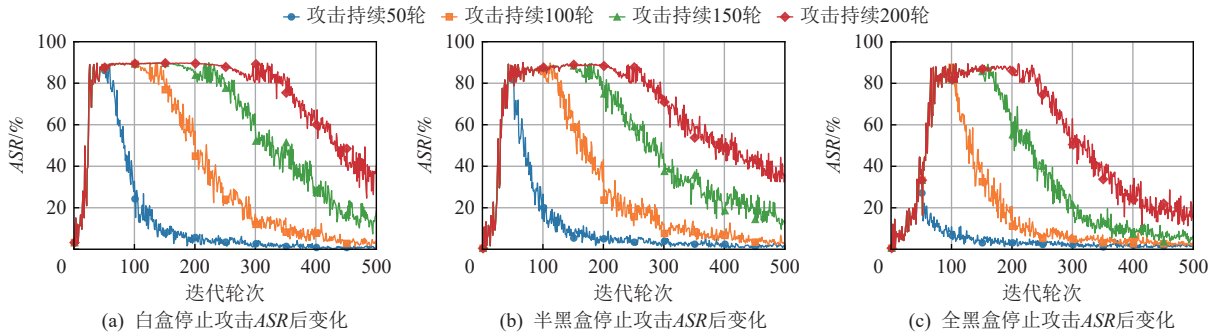


Fig. 8 *ASR* comparative analysis of AGABA framework component in ablation experiments

图 8 AGABA 框架组件在消融实验中的 *ASR* 对比分析



(a) 白盒停止攻击*ASR*后变化 (b) 半黑盒停止攻击*ASR*后变化 (c) 全黑盒停止攻击*ASR*后变化

Fig. 9 Comparative analysis of rounds forgetting after attack cessation

图 9 停止攻击后的轮次遗忘性对比分析

略,其累积的梯度方向是随机的,与主任务梯度可能存在较大相关性($\cos\theta$值较大),停止攻击后,主任务更新会直接覆盖或抵消原有的后门扰动,导致快速遗忘。半黑盒攻击则介于两者之间,基于局部损失的方向调整存在较大噪声和偏差,导致其嵌入的正交性劣于白盒,但优于全黑盒。

图 10 表示后门残存率数据,揭示了攻击持续时间与信息访问级别对后门持久性的影响,其中残存率如式(23)所示,*ASR_Final* 为最终的后门攻击成功率,*ASR_at_Stop* 为停止攻击时的后门攻击成功率,

$$persistence_score = \frac{ASR_Final}{ASR_at_Stop}。\tag{23}$$

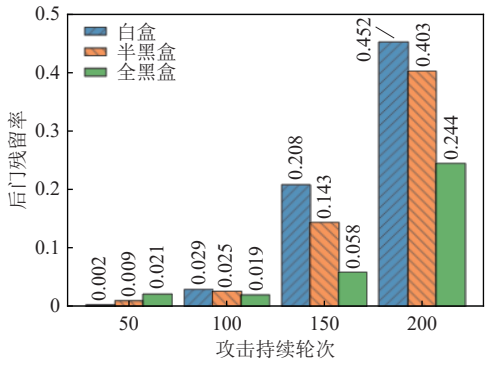


Fig. 10 Comparison of backdoor persistence rates

图 10 后门残存率对比

残存率在此定义为攻击停止后,后门功能在经过后续正常训练后得以保留的相对强度,其高低直接反映了后门嵌入在模型参数空间中的结构稳定性和抗干扰能力。持续时间较短的情况(50 轮),3 种场景因攻击的累积基数过小,残存率均处于极低水平,而白盒自适应机制在攻击初期基于尚不稳定的反馈信号进行动态调整,可能导致攻击策略在初始阶段出现振荡,从而降低了梯度累积效率与一致性。随着攻击持续时间延长至 100 轮和 200 轮,3 种场景的残存率均呈现增长,即使攻击轮次增加,黑盒场景方向优化缺陷,后门嵌入始终停留在质量较低的层次,留存率增长最为缓慢,信息反馈对于梯度累积方向的优化起决定性作用。

4.3 局限性

AGABA 框架在受控验证环境中虽验证了基于信息反馈的梯度累积攻击的有效性,但其核心机制在面向大规模复杂模型时存在理论层面的局限。在复杂模型的高维参数空间中,实现后门梯度G_{accum}与主任务梯度G_{main}的方向正交难度呈指数级增加。大规模神经网络的损失景观高度非凸,主任务梯度方向动态多变,使得基于有限轮次反馈(如全局 *ASR*)的 AST 和 MGA 策略难以准确捕捉并收敛到一个稳定的、具有高正交性的攻击方向。在大规模联邦场景中,数据高度异构,导致全局 *ASR* 或本地损失等反馈信号噪声增大、延迟升高,严重干扰 AST 与 MGA 的决策精度。同时,MGA 中的参数重要性计算需要对模型进行前向扰动,其计算开销随模型参数量线性增长,在大模型上不具有可行性。因此,AGABA 当前的设计更适用于低维可控环境,未来需进一步探索能够适应高维动态优化、信号噪声鲁棒且计算高效的新型自适应攻击框架。

5 结 论

本文提出一种面向联邦学习场景的自适应梯度累积后门攻击框架(AGABA),通过整合 AST 和 MGA 机制,使恶意更新在不牺牲隐蔽性的前提下保持高效性。AST 模块采用多因子耦合的动态透明度策略与分布式子块随机组合,实现了对触发器显隐程度控制,MGA 采用 3 阶段攻击策略将攻击载荷渐进式地注入到多轮训练中,有效规避了检测机制。在大部分测试场景下,AGABA 都能有效突破防御限制并保持较高的攻击成功率,现有防御机制虽提供了一定

程度的安全保障，但在面对高级后门攻击时仍然存在脆弱性。本工作亦明确了该框架的局限性，计算开销与反馈延迟也将成为攻击瓶颈。在未来工作中，我们将进一步研究针对分布式协同攻击的新型防御机制，构建兼顾安全与隐私保护的联邦学习系统。

作者贡献声明：张桂芬负责模型设计与分析；苏秦春负责算法设计，开展实验并撰写和修改论文；李鸿梓负责补充实验；葛丽娜提出指导意见并修改论文。

参 考 文 献

[1] Ge Lina, Li Haiao, Wang Xiao, et al. A review of secure federated learning: privacy leakage threats, protection technologies, challenges and future directions[J]. Neurocomputing, 2023, 561: 126897

[2] Li Haiao, Ge Lina, Tian Lei. Survey: Federated learning data security and privacy-preserving in edge-internet of things[J]. Artificial Intelligence Review, 2024, 57(5): 130

[3] Li Yankun, Ge Lina, Jiang Ming. Fine-grained personalized federated learning via transformer in the transformer framework[J]. Knowledge-Based Systems, 2024, 301: 112276

[4] Ge Lina, Li Yankun, Li Haiao, et al. A review of privacy-preserving research on federated graph neural networks[J]. Neurocomputing, 2024, 558: 128166

[5] Gu Tianyu, Liu Kang, Dolan-Gavitt B, et al. Badnets: Evaluating backdooring attacks on deep neural networks[J]. IEEE Access, 2019, 7: 47230−47244

[6] Saha A, Subramanya A, Pirsiavash H. Hidden trigger backdoor attacks[C]//Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 11957−11965

[7] Sun Gan, Cong Yang, Dong Jiahua, et al. Data poisoning attacks on federated machine learning[J]. IEEE Internet of Things Journal, 2021, 9(13): 11365−11375

[8] Xie Chulin, Huang Keli, Chen P Y, et al. DBA: Distributed backdoor attacks against federated learning[C]//Proc of the 8th Int Conf on Learning Representations. Washington: ICLR, 2020: 5830–5848

[9] Liu Tao, Zhang Yuhang, Feng Zhu, et al. Beyond traditional threats: A persistent backdoor attack on federated learning[C]//Proc of the 38th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2024: 21359–21367

[10] Bagdasaryan E, Shmatikov V. Blind backdoors in deep learning models[C]//Proc of the 30th USENIX Security Symp. Berkeley, CA: USENIX Association, 2021: 1505–1521

[11] Zhang Zhengming, Panda A, Song Linyue, et al. Neurotoxin: Durable backdoors in federated learning[C]//Proc of the 39th Int Conf on Machine Learning. New York: PMLR, 2022: 26429−26446

[12] Fang Pei, Chen Jinghui. On the vulnerability of backdoor defenses for federated learning[C]//Proc of the 37th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2023: 11800–11808

[13] Lin Tao, Kong Lingjiang, Stich S U, et al. Ensemble distillation for robust model fusion in federated learning[C]//Proc of the 34th Neural Information Processing Systems. San Francisco, CA: Morgan Kaufmann, 2020: 2351–2363

[14] Wu Chen, Yang Xian, Zhu Sencun, et al. Mitigating backdoor attacks in federated learning[J]. arXiv preprint, arXiv: 2011.01767, 2020

[15] Sanh V, Wolf T, Rush A. Movement pruning: Adaptive sparsity by fine-tuning[C]//Proc of the 34th Neural Information Processing Systems. San Francisco, CA: Morgan Kaufmann, 2020: 20378−20389

[16] Wang Yongkang, Zhai Dihua, Zhan Yufei, et al. Rflbat: A robust federated learning algorithm against backdoor attack[J]. arXiv preprint, arXiv: 2201.03772, 2022

[17] Ozdayi M S, Kantarcioglu M, Gel Y R. Defending against backdoors in federated learning with robust learning rate[C]//Proc of the 35th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2021: 9268−9276

[18] Blanchard P, El Mhamdi E M, Guerraoui R, et al. Machine learning with adversaries: Byzantine tolerant gradient descent[C]//Proc of the 30th Conf on Neural Information Processing Systems. San Francisco, CA: Morgan Kaufmann, 2017: 119–129

[19] Yin Dong, Chen Yudong, Kannan R, et al. Byzantine-robust distributed learning: Towards optimal statistical rates[C]//Proc of the 35th Int Conf on Machine Learning. New York: PMLR, 2018: 5650−5659

[20] Cao Xiaoyu, Fang Minghong, Liu Jia, et al. FLtrust: Byzantine-robust federated learning via trust bootstrapping[J]. arXiv preprint, arXiv: 2012.13995, 2020

[21] Nguyen T D, Rieger P, De Viti R, et al. Flame: Taming backdoors in federated learning[C]//Proc of the 31st USENIX Security Symp. Berkeley, CA: USENIX Association, 2022: 1415–1432

[22] Guerraoui R, Rouault S. The hidden vulnerability of distributed learning in Byzantium[C]//Proc of the 35th Int Conf on Machine Learning. New York: PMLR, 2018: 3521–3530

[23] Xie Chulin, Chen Minghao, Chen Pinyu, et al. CRFL: Certifiably robust federated learning against backdoor attacks[C]//Proc of the 38th Int Conf on Machine Learning. New York: PMLR, 2021, 139: 11372–11382



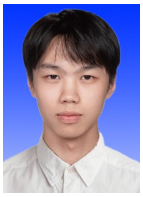
Zhang Guifen, born in 1974. Master, associate professor, master supervisor. Member of CCF. Her main research interests include information security and federated learning.

张桂芬，1974 年生。硕士，副教授，硕士生导师。CCF 会员。主要研究方向为信息安全、联邦学习。



Su Qinchun, born in 1998. Master. His main research interests include network and information security, and federated learning.

苏秦春，1998 年生。硕士。主要研究方向为网络与信息安全、联邦学习。



Li Hongzi, born in 2000. Master. His main research interests include network and information security, and federated learning.

李鸿梓, 2000 年生。硕士。主要研究方向为网络与信息安全、联邦学习。



Ge Lina, born in 1969. PhD, professor, PhD supervisor. Distinguished member of CCF. Her main research interests include artificial intelligence and information security.

葛丽娜, 1969 年生。博士, 教授, 博士生导师。CCF 杰出会员。主要研究方向为人工智能、信息安全。

《计算机研究与发展》投稿指南

征稿范围

计算机体系结构; 计算机网络与通信; 网络与信息安全; 人工智能; 软件技术; 并行与分布式计算; 图形图像、信息检索等应用技术; 其他计算机相关领域。

征稿类型

1)学术论文: 有创新学术见解的研究成果的完整论述, 对该学术领域的发展有积极意义, 或者支持我国学术生态和产业生态的重要成果。

2)综述: 对新兴的、活跃的学术研究领域或技术开发领域的评述。

3)快报短文: 对目前国内外计算机领域新思想、新观点、新技术的解读文章。

征稿要求

1)本刊只接收中文稿, 不受理英文稿。学术论文建议不超过 15 页, 综述不超过 20 页, 短文不超过 4 页。

2)要求来稿未在正式出版物上发表过, 不存在一稿多投问题。

3)作者在投稿时, 需同时向编辑部提交所有作者手写签字确认的“投稿声明”扫描版, 如不提供不予受理.对于违反投稿声明相关条款的来稿, 本刊将视情节做严肃处理, 后果自负。

4)本刊双盲评审, 作者投稿时上传系统的电子版中需要隐去作者、单位、基金、作者简介等信息, 初审通过后给编辑部邮寄的纸质版本需要包含前述署名信息。改后复审的稿件在提交修改版本以及修改说明时, 同样需符合双盲评审要求。

5)稿件评审通过后, 作者需提交单位审核盖章的“不涉密证明”和所有作者签字确认的“著作权转让声明”。

6)投稿时间在 6 个月以上的, 若没有收到编辑部发出的审理结果通知, 在书面通知编辑部并收到编辑部确认回复之后, 作者方可自行处理此稿件。

7)无论何种原因撤回投稿, 需以书面形式通知编辑部并确保所有作者知情。