基于随机断层与梯度剪裁的横向联邦学习后门防御研究

许文韬，王斌君

引用本文

许文韬，王斌君. 基于随机断层与梯度剪裁的横向联邦学习后门防御研究[J]. 计算机科学，2023，50(11)：356-363.

XU Wentao, WANG Binjun. Backdoor Defense of Horizontal Federated Learning Based on Random Cutting and GradientClipping [J]. Computer Science, 2023, 50(11): 356-363.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.221200005

基于随机断层与梯度剪裁的横向联邦学习后门防御研究

许文韬　王斌君
中国人民公安大学信息网络安全学院　北京 100038
(1625592944@qq.com)

摘　要　联邦学习解决了用户隐私与数据共享相悖之大数据困局,体现了“数据可用不可见”的理念。然而,联邦模型在训练过程中存在后门攻击的风险。攻击者通过本地训练一个包含后门任务的攻击模型,并将模型参数放大一定比例,从而实现将后门植入联邦模型中。针对横向联邦学习模型所面临的后门威胁,从博弈的视角,提出一种基于随机断层与梯度剪裁相结合的后门防御策略和技术方案:中心服务器在收到参与方提交的梯度信息后,随机确定每个参与方的神经网络层,然后将各参与方的梯度贡献分层聚合,并使用梯度阈值对梯度参数进行裁剪。梯度剪裁和随机断层可削弱个别参与方异常数据的影响力,使联邦模型在学习后门特征时陷入平缓期,长时间无法学习到后门特征,同时不影响正常任务的学习。如果中心服务器在平缓期内结束联邦学习,即可实现对后门攻击的防御。实验结果表明,该方法可以有效地防御联邦学习中潜在的后门威胁,同时保证了模型的准确性。因此,该方法可以应用于横向联邦学习场景中,为联邦学习的安全保驾护航。

关键词:横向联邦学习;后门攻击;随机断层;梯度剪裁
中图法分类号　TP391

Backdoor Defense of Horizontal Federated Learning Based on Random Cutting and Gradient Clipping

XU Wentao and WANG Binjun
College of Information and Cyber Security,People's Public Security University of China,Beijing 100038,China

Abstract　Federated learning is a methodology that solves the contradiction of big data between user privacy and data sharing, and realize the concept of “data is invisible but available”. However, the federated model is at risk of backdoor attacks in the training process. The attacker trains a attack model containing a backdoor task locally, and amplifies the model parameters by a certain proportion to implant the backdoor into the federated model. Facing the backdoor threat in the training process of the horizontal federated learning, from the perspective of the game theory, this paper proposes a backdoor defense strategy and technical proposal based on the combination of random cutting and gradient clipping. After receiving the gradient from the participants, the central server randomly chooses the neural network layer from each participant, and aggregates the gradient contributions of each participant layer by layer. Then, the central sever clips gradient parameters according to gradient threshold. Gradient clipping and random cutting can weaken the influence generated by abnormal data from minority participants. It falls into platform state when the federated model learning backdoor features, so that it keeps failing on learning backdoor features without affecting the learning process of target tasks. If the central server completes the federated learning during platform state, it can defend against backdoor attacks. Experimental results show that the proposed method can effectively defend against potential backdoor threats in fe-derated learning. At the same time, the accuracy of the model is ensured. Therefore, it can be applied in horizontal federated lear-ning scenarios, providing security protection for federated learning.

Keywords　Horizontal federated learning, Backdoor attack, Random cutting, Gradient clipping

1　引言

联邦学习[1]是一种新兴的分布式机器学习技术,最早由谷歌于2016年提出,用于体现“数据可用不可见”的理念,既保障了用户隐私,又解决了数据孤岛的问题。目前,联邦学习已成为人工智能深度学习模型与分布式训练相结合的新范式,在医学[2]、自然语言处理[3]、金融[4]等领域得到了广泛应用。

横向联邦学习[5]是最早提出的联邦学习模式,它是指以数据特征空间为导向,拓展数据样本的联邦学习模式。由于

到稿日期:2022-12-01　返修日期:2023-03-13
基金项目:国家社会科学基金重点项目(20AZD114)
This work was supported by the Key Program of National Social Science Foundation(20AZD114).
通信作者:王斌君(wangbinjun@ppsuc.edu.cn)

参与方掌握的数据特征空间基本相同,参与方之间通过横向联邦学习使所有用户数据均作为独立样本投入到联邦训练中,以实现数据样本的扩展和共享,使联邦模型有更多数据以供训练。然而,横向联邦学习作为典型的分布式机器学习技术,面临恶意参与者破坏或控制联邦模型的后门攻击等风险。

横向联邦学习中后门攻击[6-8]常见的方式有基于标签翻转的后门攻击、基于植入触发器的后门攻击、基于植入触发器后修改模型参数的后门攻击等。其中,基于植入触发器后修改模型参数的后门攻击破坏力最强。该攻击方式是恶意参与者根据全局模型参数优化选择木马触发器,并用触发器训练一个包含后门任务的攻击模型。该模型既可以完成正常的目标任务(分类或预测),又可以在触发器出现时按攻击者的意愿输出,并通过放大模型参数的方法削弱其他参与方对全局模型的贡献,从而达到攻击模型主导修改联邦模型的效果。由于触发器的存在,后门攻击变得隐蔽、难以检测。另外,鉴于联邦学习通常采用安全聚合算法[9],中心服务器无法检查参与者的参数更新,即无法检测参与者对全局模型的异常贡献,这使得防御后门攻击成为一个难点。

针对联邦学习中存在的后门攻击,研究人员提出了不同的防御策略。Sun 等[10]采用了梯度剪裁与添加噪声的方法,通过将参与者更新范数限制在阈值范围之内,并对剪裁后的全局模型添加高斯噪声,从而减轻了恶意参与者对全局模型的影响。然而,若梯度剪裁的值过于宽松,则无法防御后门攻击,但过于严苛会导致模型无法收敛。Gao 等[11]采用了限制参与方更新上传比例的聚合规则,该规则要求参与方随机选择部分参数上传至中心服务器,同时设计了新的安全聚合协议,以便服务器检查参与方是否上传了部分参数而非全部参数。然而,Li 等[12]验证了攻击者可以通过上传中毒的神经通路来完成对联邦模型的后门攻击,因此由参与方选择上传参数的方法并不能完全消除后门攻击。

本文提出了一种随机断层与梯度剪裁相结合的更新策略,在神经网络进行梯度更新时,随机断层策略要求神经网络随机抛弃一部分神经网络层的参数,只对保留参数的神经网络层进行更新,并对其进行梯度剪裁。该策略会使神经网络训练过程产生数轮的平缓期,使其特征学习变慢。

将该方案用于防御联邦学习时,中心服务器将每个参与者提交的参数与随机生成的掩码相乘,实现随机保留部分神经网络层的参数,并在梯度分层聚合结束后对梯度更新进行剪裁。由于联邦学习中正常参与者始终占多数,正常参与者之间可以实现梯度贡献的互补,联邦模型训练目标任务过程不会产生平缓期,也不会影响模型的整体收敛性;而联邦学习中攻击者占少数,后门攻击者(一般情况下,在联邦学习后期开始加入触发器)训练后门任务时更像是在训练单个神经网络模型,随机断层与梯度剪裁相结合的防御策略会使后门训练进入数轮的平缓期,从而抑制了联邦学习过程中可能存在的后门攻击,也保证了目标任务的训练。本文的主要贡献有以下 3 个方面。

1)提出了一种应对植入触发器后门攻击的神经网络随机断层更新策略,并验证了该策略会使神经网络在训练时产生平缓期。

2)验证了随机断层与梯度剪裁相结合的防御策略可以使小规模参与者的后门训练产生平缓期,而不影响大规模的目标任务训练,使后门攻击长时间无法注入联邦模型。

3)在 FEMNIST 数据集上进行了实验验证,首次成功地在非独立同分布数据集上实现了后门防御,验证了使用随机断层与梯度剪裁相结合的防御策略来防御横向联邦学习中后门攻击的有效性。

2 相关技术

2.1 联邦学习协作过程

图 1 给出了联邦学习的协作过程。中心服务器制定联邦模型,将模型及初始化参数发放给所有参与方;每轮随机从 n 个客户端中选择 m 个参与方,参与方于本地训练当前的联邦模型,并将梯度更新值经同态加密处理后上传至中心服务器。中心服务器对各参与方梯度信息进行安全聚合,并将同态加密下的聚合梯度值发放给所有参与方。各参与方解密聚合梯度后更新本地模型,完成一轮通信。如此迭代执行多轮直至联邦模型收敛。

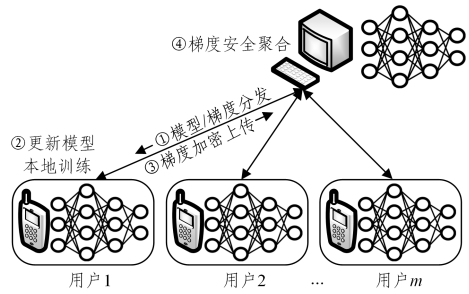

图 1 联邦学习协作过程

Fig. 1 Collaborative process of federated learning

中心服务器梯度第 $t+1$ 轮聚合过程如式(1)所示:

$$F_{\mathrm{agg}}(Ep(\Delta_{t+1}^{i}),m,\sigma)=\sigma\cdot\frac{\sum_{i=1}^{m}Ep(\Delta_{t+1}^{i})}{m}=Ep\left(\frac{\sigma}{m}\sum_{i=1}^{m}\Delta_{t+1}^{i}\right) \tag{1}$$

第 $t+1$ 轮聚合后模型最终更新情况[1]如式(2)所示:

$$\begin{aligned}G_{t+1}&=G_t+D_s\left(E_p\left(\frac{\sigma}{m}\sum_{i=1}^{m}\Delta_{t+1}^{i}\right)\right)\\&=G_t+\frac{\sigma}{m}\sum_{i=1}^{m}(L_{t+1}^{i}-G_t)\end{aligned} \tag{2}$$

其中,D_s 为同态解密算法,G_{t+1} 为第 $t+1$ 轮训练后的全局模型参数,L_{t+1}^{i} 为用户 i 第 $t+1$ 轮训练后的本地模型。

2.2 联邦学习后门攻击

本文假设后门攻击者采用基于植入触发器后修改模型参数的后门攻击[13]对联邦学习进程进行破坏,该攻击方式相比其他两种攻击方式具有更强的攻击能力,在未部署防御措施的联邦学习环境中,仅需参与一轮联邦学习通信,即可将后门注入联邦模型中。具体的攻击过程为:攻击者从中心服务器接收联邦模型,从联邦模型中选取其中一个神经网络层,并选取该层中权重之和最大的若干个神经元作为触发器的指定神经元(称为触发器神经元)。触发器神经元易被激活,且经过

神经网络训练后可以成为模型识别后门特征的重要标志。随后，攻击者使用优化的方法设计触发器，使当输入数据中存在触发器时，触发器神经元可以达到攻击者设定的目标值。优化算法如算法 1 所示。

算法 1 触发器优化算法[14]

输入：(t_i, maskx, threshold, epoch)

输出：x

```
1. f=model[:layer];
2. x=random_generator();/*随机初始化触发器 x*/;
3. while cost>threshold and e<epoch do
4.     cost=Σ(t_i−f_{n_i})²/*i=1,2,3,…*/;
5.     Δ=∂cost/∂x;
6.     Δ=Δ·mask_x;
7.     x=x−lr·Δ;/*lr 为优化学习率*/
8.     i++;
9. end while
```

其中，f 表示包含所有触发器神经元的神经网络最少前若干层，$cost$ 为损失函数，参数 $(t_i, f_{n_i})(i=1,2,\cdots)$ 分别表示第 i 个触发器神经元的目标值和当前值。

先随机初始化输入数据 x 中的触发器，将其他输入数据置零，$mask_x$ 是触发器位置掩码。每次循环将 x 输入神经网络中以计算 f，根据损失函数计算触发器神经元的目标值、当前取值及损失值。通过逐轮优化的方法调整触发器的取值来使损失值减小，直到损失值 $cost$ 小于预设损失阈值 $threshold$，或迭代次数 i 大于预设迭代次数 $epoch$。优化结束后得到触发器样式，攻击者将本地部分数据（X_{att}）嵌入触发器，并修改其数据标签（$y_{att} \rightarrow y_{new}$）。一般情况下，在分布式训练后期，将含有触发器的数据放入本地数据集中参加训练。当数据中未出现触发器时，可以以较高准确率完成目标任务；而当数据中出现触发器时，模型以高置信度按攻击者的意愿（y_{new}）进行分类或预测。最后，将攻击模型参数放大 γ 倍后，上传至中心服务器，在中心服务器进行梯度聚合时，使攻击模型主导修改全局模型。

图 2 给出了基于植入触发器后修改模型参数的后门攻击过程。

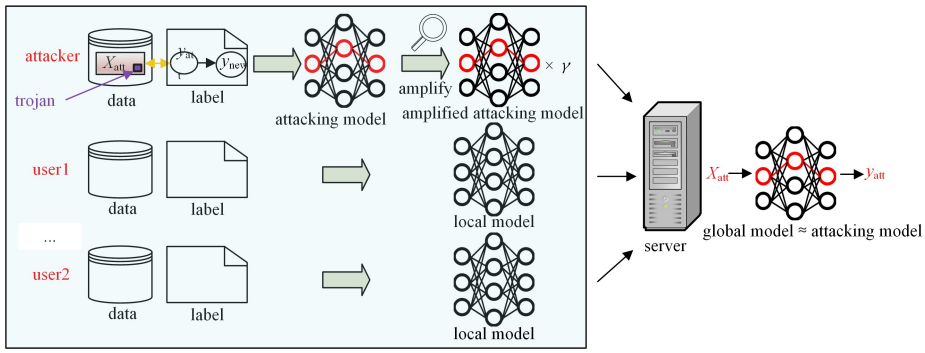

图 2 基于植入触发器后修改模型参数的后门攻击

Fig. 2 Backdoor attack based on model parameters modification after triggers implementation

由于联邦学习中后期的全局模型趋于收敛，$m-1$ 个正常参与方训练得到的本地模型与上轮全局模型相差很小，对全局模型的贡献较小，故有：

$$\sum_{i=1}^{m-1}(L_{t+1}^{i}-G_t)\approx 0 \tag{3}$$

其中，L_{t+1}^{i} 为各正常参与者第 $t+1$ 轮训练后得到的本地模型，G_t 为第 t 轮训练后的联邦模型。

攻击者可按照式(4)将攻击模型 x 的梯度变化放大 $\frac{m}{\sigma}$。

$$L_{t+1}^{att}=\frac{m}{\sigma}(x-G_t)+G_t \tag{4}$$

将式(3)和式(4)代入式(2)得式(5)：

$$G_{t+1}=G_t+\frac{\sigma}{m}\left[(L_{t+1}^{att}-G_t)+\sum_{i=1}^{m-1}(L_{t+1}^{i}-G_t)\right]=x \tag{5}$$

由此可见，中心服务器于 $t+1$ 轮聚合所有参与者的贡献后得到的结果恰好是攻击模型 x，这意味着攻击者仅需一轮通信过程即可将后门注入联邦模型中。

基于植入触发器后修改模型参数的后门攻击具有强大的攻击能力，本文选用该方法作为攻击者的后门攻击手段，并设计防御策略与之对抗，保护横向联邦学习免受后门攻击的影响。本文假设攻击者具有以下能力：

1)攻击者完全控制某个客户端及其数据集。

2)攻击者可以控制本地训练过程、模型参数以及提交的训练结果。

3)攻击者于联邦学习后期对联邦模型进行基于植入触发器后修改模型参数的后门攻击，且攻击者已知缩放因子取值，可在一轮联邦学习通信内将后门任务注入未部署防御策略的联邦模型中。

2.3 梯度剪裁

梯度剪裁[15]通常被用来解决神经网络训练过程中可能存在的梯度爆炸问题，当然梯度剪裁策略也可以减缓后门特征的学习进程，在一定程度上缓解后门攻击所带来的影响。

Sun 等[10]所采用的梯度剪裁技术为 L2 范数（欧几里得范数）剪裁，如式(6)所示：

$$\Delta t=\Delta t/\max\left(1,\frac{\|\Delta t\|_2}{c}\right) \tag{6}$$

其中，Δt 为 t 轮聚合梯度值，c 为剪裁阈值。该方法适用于非加密的横向联邦学习。当用户提交的梯度 Δt 的 L2 范数值大于预定剪裁值 c 时，中心服务器会对梯度进行剪裁，将梯度整体缩小一定比例。然而，攻击者可以通过在本地损失函数中添加 Δt 损失因素，从而可能绕开中心服务器的剪裁策略[8]。

本文则采取了另一种剪裁方式——个体梯度剪裁（Individual Gradient Clipping，IGC），即对聚合后梯度更新中的

每一项参数进行逐一剪裁,将其均限制在阈值 c 以内,如式(7)所示:

$$\Delta t^i=\Delta t^i/\max\left(1,\frac{|\Delta t^i|}{c}\right),i=1,2,3,\cdots \tag{7}$$

其中,Δt^i 为 t 轮聚合梯度值 Δt 的每一项权重,c 为剪裁阈值。

相比前者,个体梯度剪裁可以更好地减缓异常神经元被激活的速度,而这种剪裁方式难以被攻击者绕开,防御性能更强。

3 随机断层与梯度剪裁相结合的防御策略

3.1 随机断层与梯度剪裁相结合的参数更新策略

本文提出了一种新的神经网络训练梯度更新方法——随机断层策略,并将该策略与个体梯度剪裁策略相结合,提出了随机断层与梯度剪裁结合的参数更新策略(random Cutting and gradient Clipping,CAC)。随机断层策略指在神经网络的训练过程中,完成一轮训练后,抛弃一部分神经网络层的梯度更新,而其他神经网络层参数正常更新。

图 3 以 4 层神经网络为例,应用 CAC 更新策略的神经网络梯度更新过程。首先,将数据输入神经网络中进行前向传播,根据损失进行反向传播生成梯度更新值(Δt_1,Δt_2,Δt_3)。掩码生成器每轮随机生成 0-1 掩码,如(1,0,1),其中,掩码中 0 与 1 按照 η=sum(0)/sum(all)进行分配。将梯度更新值与掩码相乘得到本轮可更新参数的梯度值,如(Δt_1,0,Δt_3)。然后,使用个体梯度剪裁策略对梯度值超出阈值范围的所有参数按照式(7)进行剪裁。

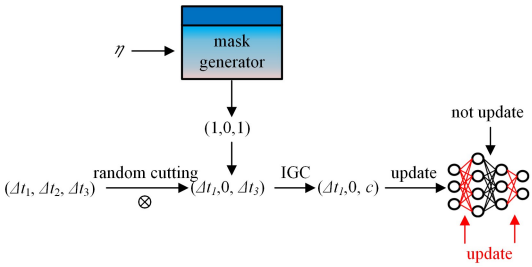

图 3 应用 CAC 策略的神经网络

Fig. 3 Neural network using CAC strategy

3.2 集中式机器学习 CAC 策略

在集中式机器学习环境下应用 CAC 策略时,掩码的存在导致一些神经网络层的参数得到的更新次数较少,且个体梯度剪裁也使得参数更新进程变慢。神经网络模型在训练初期的更新变慢,目标任务的准确率始终处于较低水平的现象称为平缓期。图 4 给出了采用正常更新策略、采取个体梯度剪裁更新策略与采取 CAC 策略的神经网络模型的准确率上升曲线。

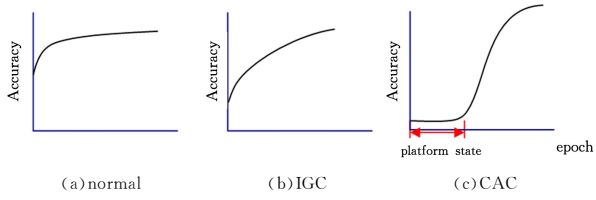

图 4 不同更新策略的收敛曲线

Fig. 4 Convergence curves of different updating strategies

个体梯度剪裁策略限制了梯度的单次更新上限。当其他条件相同时,采取个体梯度剪裁策略的神经网络模型进行多轮更新的梯度更新量与采取正常更新策略的神经网络的一轮梯度更新量相同,个体梯度剪裁策略使神经网络准确率上升速率减缓,但未产生平缓期;而应用了 CAC 策略的神经网络模型在训练初期会产生图 4(c)所示的平缓期。这是由于随机断层策略导致一些神经网络层在训练初期始终无法得到更新,前置神经网络层所学习到的特征无法向后置层传递,从而导致模型在目标任务上长时间表现不佳。同时,梯度更新受到个体梯度剪裁的限制,使模型学习特征的速度不佳,因此,出现了平缓期的现象。平缓期的长短与 η 相关,η 越大神经网络更新的层越少,平缓期较长;反之平缓期较短。

3.3 联邦学习 CAC 策略与分层平均聚合

梯度剪裁策略是联邦学习中常见的后门防御策略。在联邦学习环境下,少数攻击者对联邦模型进行后门任务攻击训练的情况,与集中式机器学习环境下相似。攻击者所提交的异常梯度值(该值可直接将后门植入无防御策略的联邦模型中)被梯度剪裁策略截断,缓解后门攻击的效果。然而,梯度剪裁策略仍保留了攻击者的部分贡献,当攻击者连续多轮进行后门攻击后,其攻击率将呈现如图 4(b)的上升趋势,联邦模型仍会在数轮内被注入后门任务。并且,梯度剪裁适值区间较小,实用性较差。

本文将 CAC 策略应用到联邦学习防御的核心思想是:中心服务器仍要求每个参与方提交单次训练更新的全部梯度信息,在接收参与方更新的梯度值时,掩码生成器为每轮每个参与者随机生成不同的掩码,随机保留每个参与者一部分神经网络层的参数,聚合时根据每一个神经网络层提交贡献的用户数进行分层平均聚合。然后,对聚合后的梯度值执行个体梯度剪裁。其算法如算法 2 所示。

算法 2 联邦学习 CAC 策略算法(中心服务器端)

输入:联邦模型结构,e,η

输出:收敛模型

1. 随机初始化联邦模型参数 θ_0
2. for round←1 to e do
3. 随机选取本轮参与方集合 index;
4. for m in index do
5. $E_p(\Delta m)$=client_update($global_{model}$,m);
6. $\mathbf{mask}^m$=get_mask(m,η);
7. $\Delta_t^m=E_p(\Delta m)\cdot\mathbf{mask}^m$;
8. end for
9. $\Delta t=\sigma\cdot\frac{\sum\Delta_t^m}{\sum\mathbf{mask}^m}=E_p\left(\sigma\cdot\frac{\sum(\Delta m\cdot\mathbf{mask}^m)}{\sum\mathbf{mask}^m}\right)$
10. Δt=individual_gradient_clipping(Δt,c);
11. distribute(Δt);
12. end for

其中,*client_update*()为参与方本地模型训练;Δm 为第 t 轮参与方 m 的梯度值;*get_mask*()为掩码生成函数,$\boldsymbol{mask}^m$ 为中心服务器为参与方 m 生成的随机掩码矩阵,用于更新相关的神经网络层;Δ_t^m 为经过断层处理后第 t 轮参与方 m 的梯度贡献;Δt 为中心服务器聚合后的梯度;σ 为全局学习率。

首先,中心服务器初始化联邦模型参数 θ_0 并同联邦模型

一起分发给所有参与者。其次，开始多轮的联邦学习：中心服务器按照采样率 r，随机选取每轮联邦学习的参与者集合 $index$；参与方执行本地模型训练后将梯度值同态加密后上传至中心服务器，中心服务器为每个参与方 k 按照 η 值随机选取神经网络层更新 0-1 掩码矩阵 $\boldsymbol{mask}^k$，并在同态加密下计算 $\Delta m \cdot \boldsymbol{mask}^m$。然后，中心服务器根据该层参与更新的用户数计算分层梯度平均聚合。最后，中心服务器执行阈值为 c 的个体梯度剪裁，并向所有参与方分发最后的梯度更新值。

在集中式机器学习环境下，CAC 策略会使模型训练初期产生平缓期，但在联邦学习环境下，CAC 策略为每个参与者分别随机选取神经网络层，各参与者之间的梯度贡献是互补的。因此，联邦模型不会像集中式机器学习模型那样，在训练初期存在神经网络层参数得不到更新而产生平缓期的情况，这是由联邦学习中正常参与者占多数所决定的。联邦学习中的恶意参与者为少数，其后门攻击更像是在部署了 CAC 策略的集中式机器学习环境下的后门攻击。由于没有其他参与者的梯度参数的互补，后门特征于断层处丢失，无法向后置神经网络层传递，从而产生数轮只针对后门任务训练的平缓期。如果中心服务器在后门攻击的平缓期内结束联邦学习，即可实现对后门攻击的防御。因此，CAC 防御策略不仅减缓了后门攻击率上升的速率，也从根本上抑制了后门攻击可能对联邦模型造成的破坏。

后门攻击者所提交的异常梯度值经过个体梯度剪裁与随机断层的双重抑制后，每次攻击者仅能对联邦模型进行微小改动，其影响力大幅降低。同时，后门攻击者受到采样率的影响，不能保证每轮都参与到联邦学习进程中。这不但使后门攻击进程变得冗长，还导致后门攻击受到灾难性遗忘的影响，即攻击者对联邦模型执行的有关后门任务的异常微小改动可能在其所未参与的联邦学习轮次中更快地被遗忘，更加有利于联邦模型实现后门防御。采取 CAC 防御策略的联邦学习针对后门训练的平缓期长短还与 η 相关，η 通常存在很大的适值区间，在适值区间内，η 取值越大，对后门防御的效果越好。

对算法 2 的性能进行分析，探究 CAC 防御策略的应用性。从时间复杂度角度分析，与标准的联邦学习环境（无防御策略、采取全局平均聚合）相比，算法 2 加入了随机断层、个体梯度剪裁和分层聚合技术。与后两者相比，随机断层是神经网络层级别的运算，在联邦学习应用中，其可穿插于等待其他参与方提交梯度贡献的时间间隙实现。而对运算时间贡献更大的个体梯度剪裁与分层聚合技术均是神经网络参数量级的运算。全局平均聚合算法可解释为 $\eta=0$ 的分层聚合，即保留每个参与者的全部神经网络层参数。因此，部署 CAC 防御策略的联邦学习与标准的联邦学习环境具有相同的时间复杂度。

从空间复杂度角度分析，中心服务器需储存各参与方梯度贡献，同时，部署 CAC 防御策略的中心服务器还需存储针对各参与方所生成的掩码。掩码中元素的个数为神经网络层数的总和，而神经网络中参数的数量远多于神经网络层数，故储存掩码所需的空间远小于储存梯度参数所需的空间。因此，部署 CAC 防御策略的联邦学习与标准联邦学习环境具有相同的空间复杂度。

综上所述，CAC 策略的部署，可实现联邦学习防御后门攻击，但并未显著增加时间与空间的开销，具有实际应用价值。

4 实验分析

将 CAC 策略部署在集中式机器学习环境下，验证了该更新策略会使模型训练初期产生平缓期；将 CAC 防御策略与分层平均聚合算法部署在联邦学习环境下，验证了该防御策略不会使正常任务的训练产生平缓期，而后门任务的训练会产生平缓期。从而证明了该防御方法比仅采取个体梯度剪裁的防御策略能够更好地防御联邦学习中的后门攻击，同时验证了该防御策略可以拓展个体梯度剪裁的适值区间。

4.1 实验设置

4.1.1 数据集

本文使用 FEMNIST[16] 数据集的一个子集进行实验验证，具体信息如表 1 所列。FEMNIST 数据集是由 Sebastian 等收集的、包括 3550 个用户共 805263 个手写体数字的图片数据集，包含了数字和大小写字母共 62 种数据类别。FEMNIST 数据集中每个用户书写风格不尽相同，故数据集属于 non-i. i. d[17] 性质。

表 1 数据集基本信息

Table 1 Basic information of datasets

Dataset	Total samples	Picture size	Category	Test data
FEMNIST	156000	1 * 28 * 28	62	18124

攻击者按照算法 1 自适应地设计合适的触发器，并将触发器随机注入攻击者本地数据集的一半数据中，修改其标签为 0，用于训练后门任务，其余数据用于训练正常的目标分类任务。

4.1.2 模型与参数设置

本文选用了 7 个卷积层、3 个最大池化层、3 个全连接层作为神经网络模型（下文简称 model），如图 5 所示。实验将分别在集中式机器学习与联邦学习环境下进行。

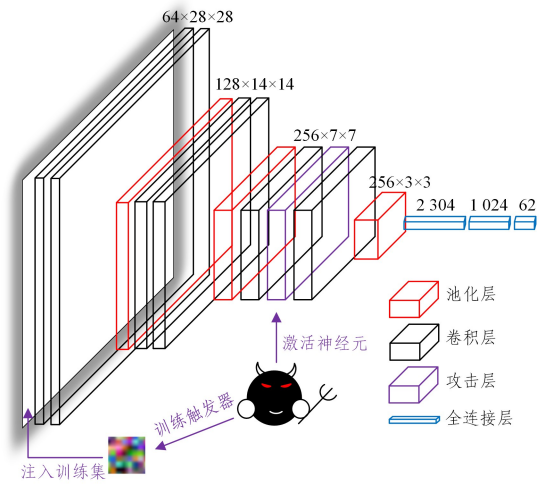

图 5 实验模型与攻击行为

Fig. 5 Experimental model and attacking behaviors

集中式机器学习环境中模型训练运行 100 轮，学习率 $\sigma=0.0003$，可选更新策略有：

1）常规更新策略。

2）个体梯度剪裁更新策略，$clip=0.0005$。

3）随机断层更新策略，$\eta=0.5$。

4）CAC 更新策略，$clip=0.0005$，$\eta=0.5$。

联邦学习实验参数信息如表 2 所列。参与者中包含 1 名后门攻击者，在联邦模型中选取 3 个易激活神经元，激活目标值为 3，并按照算法 1 获得自适应触发器，并于联邦学习第 20 轮（模型已趋于收敛）开始对联邦模型进行攻击。为了保持一致性，攻击者从 20 轮起于偶数轮参与训练，而奇数轮不参与训练。

表 2 联邦学习参数信息

Table 2 Parameter information of federated learning

Number of participants/m	Sampling rate/r	Global learning rate/σ	Scale factor/($\gamma=m\cdot r/\sigma$)
50	0.5	0.8	31.25

中心服务器可采取的防御策略有：

1）无防御策略。

2）个体梯度剪裁防御策略。个体梯度剪裁的适值区间为[0.0002，0.0005]。

3）CAC 防御策略。$\eta\in\{0.5,0.9\}$，个体梯度剪裁的适值区间可拓展为[0.0002，0.006]。

4.1.3 评价标准

本文采用以下指标来评价实验效果。

1）后门攻击率（Attack Rate）：当测试集数据嵌入后门触发器时，模型将数据归为攻击者指定类别的数据总数占全部测试集数据量的比例。

2）主任务准确率（Accuracy）：当测试集数据不包含后门触发器时，模型将数据正确分类的比例。

4.2 集中式机器学习 CAC 策略

在集中式机器学习环境，采取正常更新策略、个体梯度剪裁更新策略、随机断层更新策略以及 CAC 策略分别对 model 进行训练，其主任务准确率上升情况如图 6 所示。

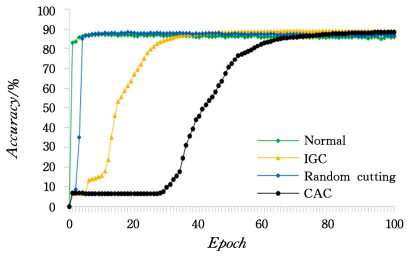

图 6 集中式机器学习环境应用不同更新策略

Fig. 6 Various updating strategies under centralized machine learning environment

从图 6 可知，常规更新策略的模型训练可在第 3 轮后趋于收敛；单独应用随机断层策略不会显著影响模型的收敛速度；个体梯度剪裁策略的应用显著延缓了模型收敛的速度，这是因为每轮的梯度更新量受到梯度剪裁的限制，所以需要更多的轮次来完成；应用 CAC 策略可以大幅减缓模型收敛的速度，同时在模型训练初期产生了明显的平缓期。

4.3 联邦学习 CAC 策略

在联邦学习环境下部署 CAC 策略（$\eta=0.5$，$clip=0.0005$），并将其收敛速度与集中式机器学习环境下部署 CAC 策略（$\eta=0.5$，$clip=0.0005$，无攻击者）的模型收敛速度进行对比，如图 7 所示。

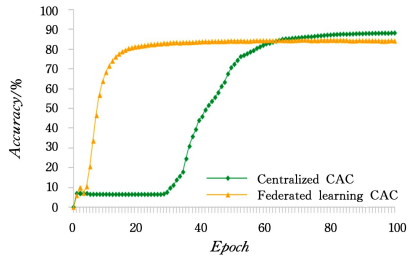

图 7 CAC 策略部署在不同机器学习环境

Fig. 7 CAC policies deployed in different machine learning environments

在联邦学习中部署 CAC 策略时，模型在训练初期并不会产生平缓期，因此，部署 CAC 策略的联邦学习正常训练过程不会受到随机断层策略的影响，这是由于中心服务器虽然使每轮每个参与者的梯度更新产生断层，但多数的正常参与者之间的梯度贡献可以形成互补，其每个神经网络层在每轮联邦学习中仍可得到更新。

因此，CAC 策略的部署，不会使联邦学习在目标任务的训练上产生平缓期而影响模型的收敛速率。

4.4 有效性分析

本文通过在联邦学习中部署后门攻击者，并分别部署无防御策略、仅个体梯度剪裁防御策略和 CAC 防御策略，统计采取不同防御策略时主任务准确率与后门攻击率的上升情况，从对后门攻击率的抑制与对有效剪裁域的拓展两方面验证 CAC 防御策略的有效性。

4.4.1 平缓期对后门攻击的抑制

通过实验测试，当仅采取个体梯度剪裁防御策略时，梯度剪裁值适值区间为[0.0002，0.0005]。当梯度剪裁值的选取落在适值区间内时，联邦模型能在数轮内较好地抵御后门攻击。但仅采取个体梯度剪裁策略时，后门任务的训练不会经历平缓期，后门攻击率仍可缓慢上升。当联邦模型在个体梯度剪裁策略的基础上加入随机断层策略后，可提高原梯度剪裁值适值区间内联邦模型的防御能力，利用平缓期抑制后门攻击率的上升。图 8 给出了个体梯度剪裁值为 0.0005 时，采取不同防御策略主任务准确率与后门攻击率随联邦学习轮次的变化情况。

由图 8 中的准确率上升曲线可知，个体梯度剪裁是减缓模型收敛速度的主要原因，而随机断层策略的加入并不影响模型整体收敛性，也不会降低模型收敛的速度。而由图 8 中的后门攻击率上升曲线可知，当未部署防御策略时，后门攻击者可在 1－2 轮内将后门任务注入联邦模型中；适值区间内的剪裁值对缓解后门攻击有所帮助，但后门攻击率随攻击轮次的增多仍可在数轮后缓慢上升；加入随机断层策略后，后门攻击率更低，防御性能更好，且 $\eta=0.9$ 时防御性能最好。

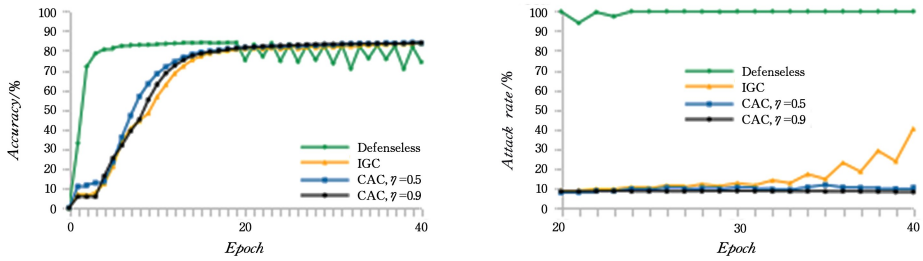

图 8 适值剪裁值下不同防御策略的性能

Fig. 8 Performance of different defensive strategies with suitable clipping

4.4.2 拓展梯度剪裁域

梯度剪裁防御策略在应对后门攻击时存在着较小的适值剪裁域。较小的剪裁值使模型长期无法收敛；剪裁值取值较大时为无效剪裁，会使后门任务仍可以在短短数轮内注入联邦模型。根据实验测试，本实验环境下联邦学习个体梯度剪裁适值区间为[0.000 2，0.000 5]。剪裁适值区间较小，使得仅采用个体梯度剪裁策略实现后门防御变得困难，而当 CAC 防御策略防御后门攻击时，可以拓展剪裁域的空间，提高了该梯度剪裁策略的实用性。

图 9 给出了当剪裁值取值处于原适值区间外，不同防御策略主任务准确率与后门攻击率随联邦学习轮次增多的变化情况。

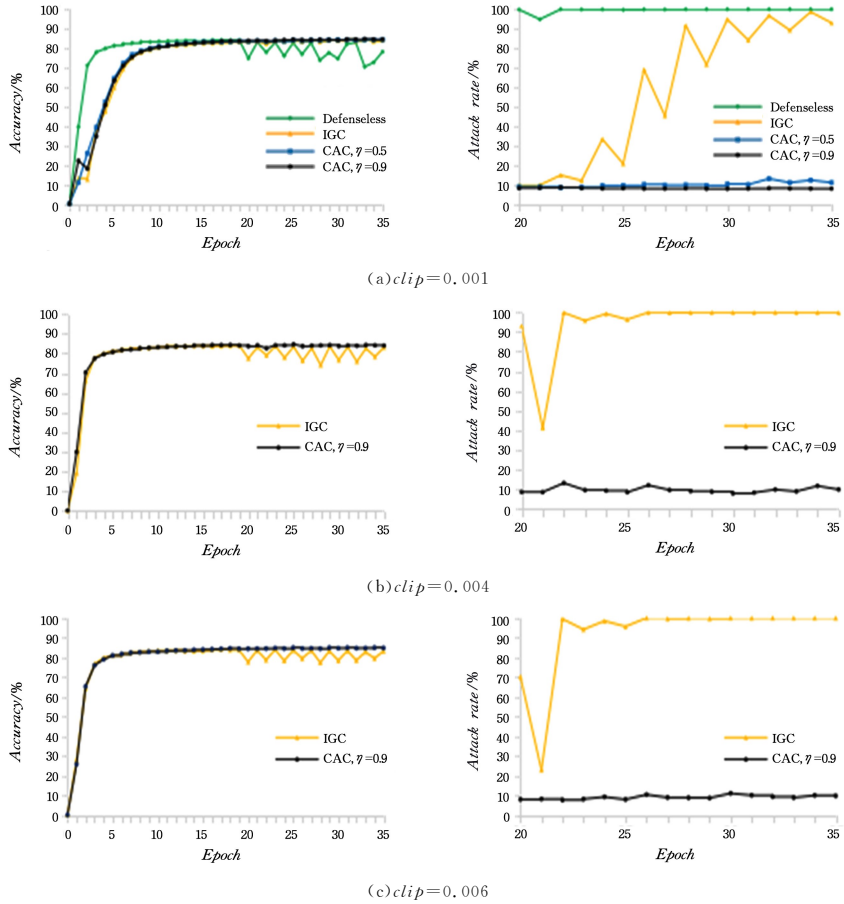

图 9 原非适值剪裁值下不同防御策略的性能

Fig. 9 Performance of different defensive strategies with original unsuitable clipping value

当个体梯度剪裁取值为 0.001 时，超出了原适值区间，攻击者攻击仅部署个体梯度剪裁($clip=0.001$)防御策略时后门攻击率逐轮上升，因此较高的剪裁值无法防御后门攻击，属无效剪裁，但攻击者在 15 轮内无法攻破部署 CAC 防御策略的联邦模型；$\eta=0.9$ 时防御性能更好，且未影响模型收敛性与主任务准确率。

当个体梯度剪裁取值为 0.004 及 0.006 时，远超出原适值区间，仅采取个体梯度剪裁策略进行防御无法有效抵御后门攻击，后门攻击者可在 1－2 轮内将后门任务注入联邦模型，此时梯度剪裁为无效剪裁；加入随机断层策略后，即便个体梯度剪裁值选取过大，$\eta=0.9$ 的 CAC 防御策略仍能有效抵御后门攻击，且未影响模型收敛性与主任务准确率。

综上所述，CAC 防御策略的应用，可以拓展个体梯度剪裁值选取的上限，在本实验中，剪裁域可拓展 1 个数量级以上，使原本无效的剪裁变为有效剪裁，大大提高了防御策略的实用性。

结束语 后门攻击是联邦学习中常见的安全威胁，具有很强的隐蔽性和破坏性，而现有的后门防御策略在抵御后门攻击时表现不佳。因此，本文提出了一种基于随机断层与梯度剪裁相结合的防御策略，该策略在不影响主任务准确率的同时，可以最大程度地限制后门攻击行为，为联邦学习中后门防御提供了新方法。

同时，将随机断层与梯度剪裁相结合的防御策略部署在用户端可以防御潜在的推理攻击行为，如深度梯度泄露[18]。梯度信息是由参与者使用本地数据经过计算得到的，若联邦学习中的恶意参与者截获用户上传的梯度信息，用户原始数据便存在泄露的风险。个体梯度剪裁策略与随机断层策略均可对梯度信息进行非线性变换，使梯度信息不再具有还原原始数据的能力。因此，基于随机断层与梯度剪裁相结合的防御策略也可为联邦学习中推理攻击的防御提供新思路。

参 考 文 献

[1] MCMAHAN B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data[C]// Artificial Intelligence and Statistics. Florida: PMLR, 2017: 1273-1282.

[2] XU J, GLICKSBERG B S, SU C, et al. Federated learning for healthcare informatics[J]. Journal of Healthcare Informatics Research, 2021, 5(1): 1-19.

[3] LIN B Y, HE C, ZENG Z, et al. Fednlp: Benchmarking federated learning methods for natural language processing tasks[C]// Findings of the Association for Computational Linguistics: NAACL 2022. Stroudsburg: ACL, 2022: 157-175.

[4] BYRD D, POLYCHRONIADOU A. Differentially private secure multi-party computation for federated learning in financial applications[C]// Proceedings of the First ACM International Conference on AI in Finance. New York: ACM, 2020: 1-9.

[5] KAIROUZ P, MCMAHAN H B, AVENT B, et al. Advances and open problems in federated learning[J]. Foundations and Trends in Machine Learning, 2021, 14(1/2): 1-210.

[6] TOLPEGIN V, TRUEX S, GURSOY M E, et al. Data poisoning attacks against federated learning systems[C]// European Symposium on Research in Computer Security. New York: Springer, 2020: 480-501.

[7] WANG H, SREENIVASAN K, RAJPUT S, et al. Attack of the tails: Yes, you really can backdoor federated learning[J]. Advances in Neural Information Processing Systems, 2020, 33: 16070-16084.

[8] GONG X, CHEN Y, HUANG H, et al. Coordinated Backdoor Attacks against Federated Learning with Model-Dependent Triggers[J]. IEEE Network, 2022, 36(1): 84-90.

[9] BONAWITZ K, IVANOV V, KREUTER B, et al. Practical secure aggregation for privacy-preserving machine learning[C]// Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2017: 1175-1191.

[10] SUN Z, KAIROUZ P, SURESH A T, et al. Can you really backdoor federated learning? [J]. arXiv: 1911.07963, 2019.

[11] GAO J, ZHANG B, GUO X, et al. Secure Partial Aggregation: Making Federated Learning More Robust for Industry 4.0 Applications[J]. IEEE Transactions on Industrial Informatics, 2022, 18(9): 6340-6348.

[12] LI S H, ZHENG H B, CHEN J Y, et al. Neural Path Poisoning Attack Method for Federated Learning[J]. Journal of Chinese Computer Systems, 2023, 44(7): 1578-1585.

[13] BAGDASARYAN E, VEIT A, HUA Y, et al. How to backdoor federated learning[C]// International Conference on Artificial Intelligence and Statistics. New York: PMLR, 2020: 2938-2948.

[14] LIU Y, MA S, AAFER Y, et al. Trojaning attack on neural networks[C]// 25th Annual Network and Distributed System Security Symposium. California: The Internet Society, 2018: 1-11.

[15] ZHANG J, HE T, SRA S, et al. Why gradient clipping accelerates training: A theoretical justification for adaptivity[J]. arXiv: 1905.11881, 2019.

[16] CALDAS S, DUDDU S M K, WU P, et al. Leaf: A benchmark for federated settings[J]. arXiv: 1812.01097, 2018.

[17] LI Q, DIAO Y, CHEN Q, et al. Federated learning on non-iid data silos: An experimental study[C]// 2022 IEEE 38th International Conference on Data Engineering (ICDE). New York: IEEE, 2022: 965-978.

[18] ZHU L, HAN S. Deep leakage from gradients[C]// Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems. New York: Curran Associates Inc, 2019: 14747-14756.

XU Wentao, born in 1999, postgraduate. His main research interests include federated learning and backdoor attack.

WANG Binjun, born in 1962, Ph.D, professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include network security and law enforcement.

（责任编辑：杨雪敏）