异步联邦学习中隔代模型泄露攻击及防治方法

胡智尧，于 淼，田开元
（军事科学院战争研究院，北京，100091）

摘要 联邦学习已成为数据孤岛背景下知识共享的成功方案。随着梯度逆向推理等新式攻击手段的问世，联邦学习的安全性再度面临新挑战。针对联邦学习可能存在参与者恶意窃取其它客户端梯度信息的风险，提出一种异步联邦学习框架下的隔代模型泄露攻击方式：利用中心服务器“接收则聚合”的特点，多名恶意客户端可按照特定更新顺序，通过隔代版本的全局模型差异逆向计算其他客户端的模型更新数据，从而窃取对方的模型。针对此问题，提出基于α-滑动平均的随机聚合算法。首先，中心服务器每次收到客户端的模型更新后，将其与从最近α次聚合中随机选出的全局模型进行聚合，打乱客户端的更新顺序；其次，随着全局迭代次数增加，中心服务器对最近α次聚合的全局模型进行滑动平均，计算出最终全局模型。实验结果表明，与异步联邦学习方法相比，FedAlpha方法有效降低隔代模型泄露攻击的可能性。

关键词 异步联邦学习安全；逆向推理攻击；随机聚合；滑动平均；隔代模型泄露攻击
DOI 10.3969/j.issn.2097-1915.2024.05.016
中图分类号 TP391.9 **文献标志码** A **文章编号** 2097-1915(2024)05-0121-07

An Attacking and Prevention Method of Inter-Generational Model Leakage in Asynchronous Federated Learning

HU Zhiyao, YU Miao, TIAN Kaiyuan
(Institute of War Studies, Academy of Military Sciences, Beijing 100091, China)

Abstract Federated learning is a successful solution for shared knowledge in the context of data islands. However, with the advent of new attacks such as gradient reverse reasoning, the security of federated learning is faced with a new challenges again. In the federated learning, an inter-generational model leakage problem under the asynchronous federated learning framework is proposed aimed at the problem that participants maliciously steal gradient information from other clients by any possibility. By utilizing the characteristics of central server receiving then aggregating, multiple malicious clients can reversely compute other clients' model update data through inter-generational versions of the global model in a specific update order. In view of this problem, a random aggregation algorithm based on α moving average is proposed. Firstly, the model update being received each time, the central server is to aggregate it with the global model randomly selected from the latest α aggregations, and shuffle the clients' update order through the randomness of the aggregation. Secondly, as the number of global iterations increases, the central server performs a moving average on the global model of the latest aggregation to calculate the final

收稿日期：2023-12-07
基金项目：国家自然科学基金(62202491，62402519)
作者简介：胡智尧(1992—)，男，贵州贵阳人，助理研究员，博士，研究方向为数据安全、科技安全战略。E-mail：huzhiyao92@yeah.net

引用格式：*胡智尧，于淼，田开元. 异步联邦学习中隔代模型泄露攻击及防治方法[J]. 空军工程大学学报，2024，25(5)：121-127. HU Zhiyao，YU Miao，TIAN Kaiyuan. An Attacking and Prevention Method of Inter-Generational Model Leakage in Asynchronous Federated Learning[J]. Journal of Air Force Engineering University，2024，25(5)：121-127.*

global model. The experiment simulations show that the FedAlpha method can effectively reduce the possibility of inter-generational model leakage in comparison with the asynchronous federated learning method.

Key words asynchronous federated learning security; reverse reasoning attack; random aggregation; moving average; intergenerational gradient leakage

随着智能算法进入深度学习[1-3]时代，智能模型结构和训练算法已进入大语言模型时代。但在军队、政企、医院等单位涉及敏感数据，难以实现数据共享，数据孤岛问题已严重影响人工智能应用落地。跨单位的智能模型协同训练是一大难题。

联邦学习技术[4]是一种保护客户端数据安全的分布式学习方式，改善数据孤岛问题，提升数据利用价值。在联邦学习框架下，各客户端（参与者）从中心服务器（组织者）接收全局模型后利用本地数据对其进行训练模型；训练好的本地模型发送至中心服务器（组织者）后，由其将各本地模型聚合为统一的全局模型。由于整个训练过程中客户端的私有数据不上传至外网，从而保护本地数据安全。

联邦学习仍存在模型安全问题。文献[5]提出一种梯度逆向推理的攻击方式，采用生成对抗网络技术通过模型参数获取训练样本数据。由于中心服务器需要保管维护各客户端的模型数据，客户端的模型数据可能在中心服务器聚合的过程中被泄露。

学者将同态加密算法[6]和联邦学习进行结合，要求客户端与中心服务器进行通信时使用模型密文数据。同态加密的特性是，中心服务器无需解密就能对从客户端接收到的模型密文数据进行聚合计算，从而防止中心服务器窃取模型原始数据。但是，中心服务器的同态计算存在限制：来自不同客户端的模型密文数据遵循同一套加密算法，意味着不同客户端之间可以相互解密对方的模型数据。若恶意客户端通过非法渠道获得其他正常客户端的模型密文数据后，可直接进行破解。

在异步联邦学习框架中，不同客户端更新全局模型的速度各异。故各客户端从中心服务器得到聚合后的全局模型具有不同版本，并且更新顺序相邻的客户端有着版本连续的全局模型。这类版本连续的全局模型是按照聚合算法计算得到的。若反向利用聚合算法，可推理出更新顺序相邻的其他客户端的模型数据。特别是联邦学习客户端不一定全都是可信的，他们存在恶意攻击行为。利用上述异步联邦学习框架的漏洞，多名恶意客户端通过相互串通分享相邻版本的全局模型，可破解其他客户端模型数据。基于此思路，本文提出一种模型隔代泄露的新式联邦学习攻击行为。

针对隔代模型泄露攻击方式，本文设计了一种基于 α-滑动平均的随机聚合（FedAlpha）算法。当收到一位客户端上传的模型更新时，中心服务器将从最近 α 次聚合中随机选出一个过去的全局模型，再与该模型更新进行聚合。在每执行完 α 次聚合后，中心服务器再将最近 α 次聚合产生的全局模型进行聚合，产生最新的全局模型。实验结果表明，与异步联邦学习方法相比，FedAlpha 方法有效降低隔代模型泄露攻击的可能性，同时保证模型正常收敛。

1 相关概念

1.1 联邦学习的同步和异步聚合

联邦学习允许客户端不上传数据、直接在本地训练模型，从而避免客户端隐私泄露。传统的联邦学习框架采用同步聚合：每轮全局迭代的训练过程中，中心服务器从全部客户端里随机选出固定数量的客户端；被选中的客户端将会收到中心服务器发送的全局模型，并利用客户端设备存储的本地数据进行模型训练；本地训练结束后，客户端将训练好的本地模型发送至中心服务器；中心服务器收到客户端的本地模型后执行模型聚合操作。聚合后的模型将被用作新一轮迭代训练的全局模型，将再次被下发给随机抽选的客户端，开始新一轮的本地训练。上述过程需要一直持续，直到全局模型收敛。在同步联邦学习框架下，只有接收到所有参与本轮次训练客户端的本地模型后，中心服务器才能进行聚合。

图 1 显示了一种基于同态加密聚合的联邦学习框架。中心服务器从 10 名客户端中随机选取 3 名客户端参与训练。在传统联邦学习框架中，中心服务器采用同步聚合方式，只有在收到 3 名客户端上传的模型数据之后才进行聚合。3 位客户端的模型更新进行加权求和后再与初始模型计算，得到下一轮迭代的全局模型。下一轮全局迭代将重新选出 3 位客户端，重复上述操作直至全局模型收敛。

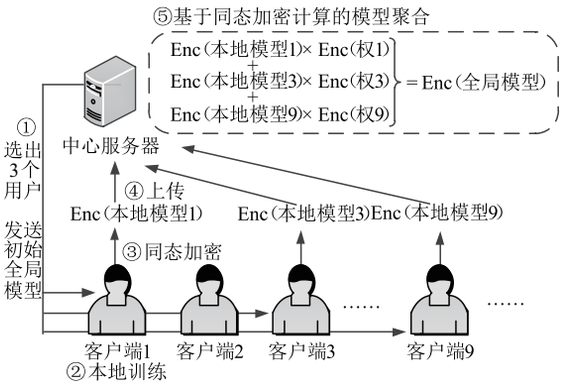

图 1　同态加密的联邦学习框架

Fig. 1　Federated learning framework using homomorphic encryption

由于网络通信条件有限，参与训练的客户端上传本地模型的时间不同。个别客户端容易因掉线等特殊情况难以及时上传本地模型，从而延误中心服务器的模型聚合。为提高联邦学习训练效率，一种中心服务器采用异步聚合方式的联邦学习框架被提出。中心服务器将全局模型发送给全部客户端后，各客户端开始并行训练；响应时间快的客户端能更快完成训练并将模型传回至中心服务器。在第 i 轮全局迭代训练中，中心服务器一旦收到一位客户端 c 上传的模型更新 U_c，立即将其与最新版的全局模型 W_G^i 进行聚合，产生下一版的全局模型 W_G^{i+1}；该全局模型将被中心服务器发送给客户端 c，由其利用本地数据开始新一轮的迭代训练。若客户端 c 上传的模型更新 U_c 不是用最新版的全局模型 W_G^i 训练得到的，而是用第 j 轮全局迭代的全局模型 W_G^j 训练得到的（$i \neq j$），则 U_c 具有陈旧性。相关研究表明[7]，直接用陈旧的本地模型进行更新，将降低全局模型的收敛速度，故在聚合前需进行陈旧系数进行惩罚。选取一个小于 1 的惩罚系数 β，将 U_c 乘以 β^{i-j} 进行衰减处理后，再执行聚合操作。

总的来看，异步联邦学习框架较传统联邦学习具有更高的训练效率，有效解决同步延误问题，得到广泛应用。

1.2　基于同态加密的联邦学习

同态加密是一种数据加密后仍能进行加乘计算的加密技术，具有同态计算性质。同态加密的数据进行加乘计算后执行解密操作得到的结果，与未加密的原始数据经相同计算后的结果相等。如图 1 所示，中心服务器收到各客户端同态加密并上传的模型参数后，无需解密、直接对客户端的模型更新数据（密文）进行加权求和。图中 Enc(·) 表示同态加密算法。数据 a 与数据 b 同态加密后满足：

$$\mathrm{Enc}(a)+\mathrm{Enc}(b)=\mathrm{Enc}(a+b)$$

$$\mathrm{Enc}(a)\mathrm{Enc}(b)=\mathrm{Enc}(ab)$$

经过聚合计算后，中心服务器将全局模型（密文）发送给下一轮训练的客户端，客户端解密后可直接进行本地训练。在上述聚合过程中，中心服务器不曾知晓客户端模型原始数据（明文），故同态加密有效防止客户端上传的模型原始数据被泄露，避免逆向推理攻击。

值得注意的是，同态加密包括对称加密和非对称加密。对称同态加密是指加密所用的密钥和解密所用密钥一致；非对称加密则不要求两者一致。在联邦学习中，非对称的同态加密更为实用。一方面，中心服务器需要将未加密的初始全局模型进行加密，再同客户端上传的模型数据（密文）进行聚合。故中心服务器只持有公钥，只能进行加密操作，但不具备私钥而无法对客户端上传的模型数据进行解密操作。另一方面，中心服务器聚合模型后，发给客户端的是全局模型的密文数据。客户端用私钥对收到的模型数据解密，再进行新一轮本地迭代训练。故客户端同时持有公私钥，既可加密又可解密。

1.3　联邦学习的隐私威胁与防治

联邦学习推理攻击是指从现有的人工智能模型中获取未被公开的信息，主要分为模型逆向攻击、属性推理攻击和成员推理攻击。其中，最常见的是模型逆向攻击和成员推理攻击。模型逆向攻击可以恢复用于训练受攻击模型的数据；属性推理攻击是窃取训练模型所使用的数据统计信息，如样本的类别分布等；成员推理攻击可以推断某一样本是否在训练数据集中。

常见的防御技术包括同态加密、差分隐私、安全多方计算。同态加密是联邦学习框架必备的安全手段之一。如前文所述，同态加密可以避免泄露客户端上传的模型原始数据，同时加密后的模型数据不妨碍中心服务器进行聚合计算操作。差分隐私技术[8]是通过向训练过程中的模型添加噪声，避免恶意客户端进行成员推理攻击。但添加噪声后，模型训练收敛速度将减慢，同时模型的准确性下降。安全多方计算[9]常用于存在多个中心服务器的联邦学习框架中。客户端将上传的模型数据拆分为多份不重复的数据块，分发给不同的中心服务器；中心服务器通过安全多方计算协议对数据块进行模型聚合；客户端从多个中心服务器接收经过聚合计算后的数据块，恢复出全局模型。

总的来看，上述 3 种防御技术的前提是事先获得对方的模型。现有防御技术侧重于以密码学的方法保护客户端模型数据，本文研究对象是异步联邦学习框架下多名恶意客户端串通窃取其他正常客户端的模型数据。上述防御技术并不能有效解决本文

提出的隔代模型泄露攻击。

2 隔代模型泄露攻击

2.1 问题分析

在异步联邦学习框架下，中心服务器一旦收到一个客户端上传的模型参数立即进行聚合，并将聚合后的模型发回至该客户端供其继续迭代训练。由于各客户端的本地训练时间、模型传输的网络速度不同，各客户端更新的是不同版本的全局模型。利用异步联邦学习的特点，多个恶意客户端之间相互合作、共享不同版本的全局模型，可计算出其他正常客户端的模型参数。假设相邻三代的全局模型的版本分别为 t，$t+1$，$t+2$，只要恶意客户端持有第 t 代、第 $t+2$ 代的梯度模型，可反推第 $t+1$ 代模型中的客户端更新数据。该攻击方式称为隔代模型泄露攻击。

以异步联邦学习框架为例，图 2 显示了多个恶意客户端隔代模型泄露攻击的过程。客户端 1、客户端 2、客户端 3 按照先后顺序更新全局模型。如果客户端 1 和客户端 3 是恶意客户端，根据全局模型的更新版本可确定 3 位客户端的更新顺序。首先，恶意客户端 3 上传本地模型 W_3 后可收到中心服务器发回的 W_G^3。利用聚合规律，客户端 3 可计算出 W_G^2。由于 W_G^2 是客户端 2 更新全局模型 W_G^1 后得到的，并且恶意客户端 1 收到过中心服务器下发的 W_G^1。根据聚合算法，客户端 1 通过 W_G^1 和 W_G^2 就能计算出客户端 2 的本地训练模型 W_2。故客户端 1 与客户端 3 相互串通能推算出 W_2。至此，正常客户端 2 的模型遭到泄露。

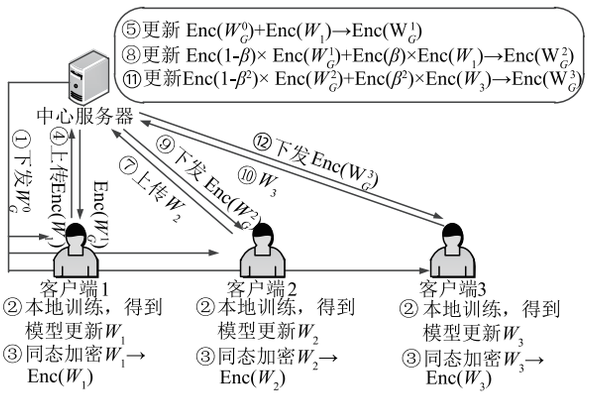

图 2 异步联邦学习框架的隔代模型泄露攻击示意图

Fig. 2 Inter-Generational model leakage illustration of an asynchronous federated learning framework

当模型隔代泄露攻击成功后，正常客户端上传的本地模型被恶意客户端窃取。但此时只能计算出正常客户端的模型梯度，还不能从中获取正常客户端的隐私数据。若恶意客户端进一步通过梯度泄露攻击[5]，利用生成式对抗网络的训练原理，通过减少梯度差异的方式，不断生成与正常客户端的训练数据，从而获取用户隐私。

由于 W_2 是客户端 2 从中心服务器收到的 W_G^0 上经过本地训练得到的，由此计算出客户端 2 的梯度更新。获得梯度更新后，恶意客户端再利用梯度泄露攻击手段[5]逆向推理出客户端 2 的本地数据。

值得注意的是，恶意客户端的目标是正常客户端的模型数据。为此，恶意客户端不必真正参与到联邦学习中，甚至不经过本地训练就直接将人为生成、具有欺骗性的虚假模型数据上传至中心服务器，以套取正常客户端的模型数据。

2.2 危害分析

隔代模型泄露攻击的充要条件是，中心服务器进行异步聚合时正常客户端的更新顺序应位于 2 名恶意客户端之间。因此，客户端的更新顺序将是隔代模型泄露攻击的关键。由于该类型攻击需要多个恶意客户端配合，恶意客户端的数量与总客户端数量的比值、联邦训练的迭代轮数分布都会影响隔代模型泄露攻击的可能性。

2.2.1 恶意客户端数量与隔代模型泄露攻击的风险正相关

中心服务器采用均匀随机采样的方式，总客户端数量设置为 100，设置客户端响应时间服从均匀分布 $U(0,100)$。实验进行 1 000 次迭代，控制恶意客户端与正常客户端的比值从 1∶9 增加至 9∶1。实验结果如图 3 所示，恶意客户端的隔代模型攻击并不总随着恶意客户端数量增加而增加。相反地，当恶意客户端数量比重超过一定阈值后，恶意客户端造成的梯度泄露次数下降。分析认为，当恶意客户端比重为 0.9 时，参与训练的客户端大部分是恶意客户端，客户端更新次序将极少包含正常客户端，故梯度泄露次数不增反减。另一方面，在相同的恶意客户端数量比重下，客户端总数越少，梯度泄露次数越多。

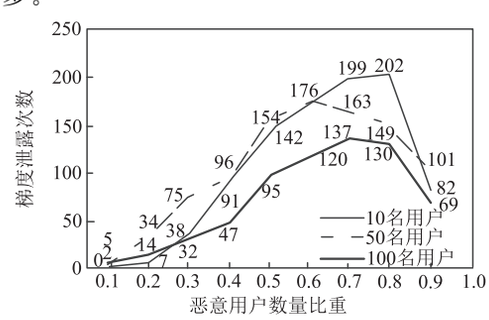

图 3 客户端数量对隔代模型泄露次数的影响

Fig. 3 Influence of the number of clients on inter-generational model leakages

2.2.2 迭代次数与隔代模型泄露风险正相关

由于各客户端的本地数据是非独立同分布的，联邦学习需要更多轮次的迭代训练才能使模型收

敛。考虑到迭代次数越多，恶意客户端进行隔代模型攻击的机会越大。实验设置 100 名客户端分别进行 500、1 000、5 000、10 000 次迭代的联邦学习，其他设置与上文相同。测试结果如图 4 所示，随着迭代轮数增加，恶意客户端成功发起梯度泄露攻击次数的可能性也增加。

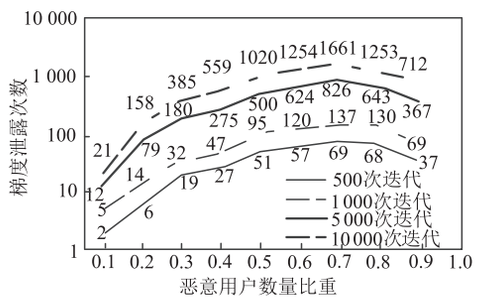

图 4　迭代次数对隔代模型泄露次数的影响

Fig. 4　Influence of the number of iterations on inter-generational model leakages

2.3　防治方法设计

解决隔代模型泄露攻击的难点是，在不确定客户端是否恶意的前提下，不改变异步联邦学习框架的聚合方式，同时实现安全的聚合。在异步联邦学习框架中，中心服务器总是将最近一次收到的客户端模型更新数据和当前最新的全局模型进行聚合。故恶意客户端利用此聚合方式可以推算出客户端的更新顺序，从而发起隔代模型泄露攻击。

针对隔代模型泄露攻击，设计一种采用 α-滑动平均随机聚合算法（FedAlpha）的异步联邦学习框架。算法主要流程如表 1 所示。

表 1　基于 α-滑动平均的 FedAlpha 算法

Tab. 1　α-moving average based on FedAlpha

基于 α-滑动平均的随机聚合算法流程
输入：参数 α，陈旧惩罚系数 β
1. 中心服务器初始化全局模型 W_G^0
2. 中心服务器将 W_G^0 发送给客户端
3. 客户端进行本地训练
4. For 第 t 次通信，$t=1,2,\cdots,T$
5. 中心服务器收到客户端 c 上传的模型密文 w 及版本数 j
6. 中心服务器从最近 α 次聚合后产生的全局模型随机选出一个全局模型，其版本 i 满足 $t-i\leqslant\alpha$ 且 $j\geqslant i$
7. 中心服务器进行聚合，$W_G^{t+1}=(1-\beta^{j-i})W_G^i+\beta^{j-i}w$
8.　If t 能被 α 整除
9.　$W_G^{t+1}=\frac{1}{\alpha}\sum_{k=0}^{\alpha-1}W_G^{t+1-k}$
10.　EndIf
11. 中心服务器将 W_G^{t+1} 发送给客户端 c
12. 客户端 c 对 W_G^{t+1} 进行本地训练
13. EndFor

该算法的创新之处包括以下 2 个方面：

1）随机聚合。当收到一位客户端上传的模型更新时，传统异步联邦学习框架下中心服务器将此更新与最新的全局模型聚合。此聚合无疑暴露各客户端的更新顺序。FedAlpha 方法采用随机聚合，从最近 α 次聚合后产生的全局模型中随机选出一个，再与该模型更新进行聚合。此操作可增加用户更新顺序的随机性，恶意客户端无法确定最近更新的客户端所使用的全局模型版本。此外，FedAlpha 方法随机聚合的是过去版本的全局模型，从而缓解模型陈旧性的负面作用。

2）滑动平均。由于 FedAlpha 方法聚合的不总是最新全局模型，训练过程的收敛速度下降。为解决此问题，FedAlpha 方法对过去训练的全局模型进行滑动平均。在每执行完 α 次聚合后，中心服务器再将最近 α 次聚合产生的全局模型进行聚合，产生最新的全局模型。值得注意的是，当 $\alpha=1$ 时，基于 α-滑动平均的随机聚合算法等价于传统的异步联邦学习方法。实验部分表明，α 取值增加有助于降低梯度泄露的可能性，并且当 α 取值较小时，模型收敛不会明显下降。总的来看，基于 α-滑动平均的随机聚合算法在模型性能可接受的下降范围内，显著提高异步联邦学习的安全性。

3　实验

实验包含 2 部分，首先测试并比较 FedAlpha 方法与传统异步联邦学习方法 FedAsync[10] 的隔代模型泄露可能性，然后测试并比较 FedAlpha 方法与 FedAsync 方法的测试准确率。

本文实验对所有方法采用相同的实验设置：联邦学习总客户端共 1 000 名，恶意客户端数量占比是 0.6，总共执行 1 000 次迭代训练，本地训练执行 10 次，batch 大小设置为 10。FedAlpha 算法的滑动平均窗口（即参数 α）分别设为 4、7、10，陈旧惩罚系数 β 设为 0.7；FedAsync 算法的陈旧惩罚系数按其论文预设参数进行设置，$\beta_t=\frac{\beta}{a(s-b)+1}$，其中 $a=10$，$b=4$，$s=j-i$ 是模型聚合时客户端模型更新和待更新全局模型的版本之差。

实验采用 2 个开源数据集，即 MNIST 和 FEMNIST。MNIST 数据集有 61 639 个训练样本和 7 396 个测试样本，用于训练有 2 个卷积层的卷积神经网络模型。FEMNIST 是重采样的联邦扩展 MNIST，包括 18 345 个训练样本和 2 136 个测试样本，用于训练 6 个卷积层的卷积神经网络模型。FedAlpha 方法和两类模型采用基于 Tensorflow 框架[11] 实现，模型参数数量分别为 50 186 和 325 578。

考虑到客户端响应时间决定了客户端上传本地模型、对全局模型进行更新的频率。若恶意客户端

响应时间较低，可频繁更新全局模型，增加隔代模型泄露的风险。为评价客户端响应时间的影响，实验采用不同的客户端响应时间分布。实验模拟 2 种重尾分布：帕累托(Pareto)分布和对数正态(logarithmic normal，LogNorm)分布下的响应时间。在 Pareto 分布中，分布形状参数和尺度参数分别设为 10。在 LogNorm 分布中，均值和标准差分别为 3 和 0.3。

3.1 隔代模型泄露可能性评价

实验指标采用隔代模型的泄露可能性，即泄露次数与总迭代训练次数之比。如表 2 所示，当响应时间分布为 LogNorm 时，FedAsync 方法的泄露可能性是 14.4%，意味着 1 000 次迭代中共发生 144 次模型隔代泄露。当设置 α 取值为 4 时，FedAlpha 方法将此情况下的泄露次数降低到 9 次。FedAlpha($\alpha=10$)方法仅发生一次梯度泄露。在 Pareto 分布下，FedAlpha($\alpha=10$)方法则无一次梯度泄露。相比而言，FedAsync 方法分别在 2 种分布下发生了 144 和 166 次隔代模型泄露。实验结果表明，FedAlpha 方法较 FedAsync 方法更好地对抗隔代模型泄露攻击。

此外，滑动平均窗口参数与隔代模型泄露可能性呈现反相关。α 取值从 4 增加到 10，梯度泄露情况减少。分析认为，α 决定了模型聚合的随机性，使得恶意客户端难以准确推测客户端的更新顺序，从而防御隔代模型泄露攻击。

表 2 隔代模型泄露可能性与准确率

Tab. 2 Possibility and accuracy of model leakages

算法	响应时间分布	数据集	泄露可能性/%	准确率/%
FedAsync	LogNorm	MNIST	14.4	87.9
		FEMNIST		85.5
	Pareto	MNIST	16.6	90.1
		FEMNIST		88.7
FedAlpha ($\alpha=4$)	LogNorm	MNIST	0.9	87.2
		FEMNIST		84.8
	Pareto	MNIST	1.5	89.3
		FEMNIST		87.9
FedAlpha ($\alpha=7$)	LogNorm	MNIST	0.5	86.4
		FEMNIST		84.1
	Pareto	MNIST	0.4	88.2
		FEMNIST		86.6
FedAlpha ($\alpha=10$)	LogNorm	MNIST	0.1	85.3
		FEMNIST		83.4
	Pareto	MNIST	0	87.5
		FEMNIST		85.7

3.2 模型收敛性评价

为评价 α-滑动平均随机聚合算法的收敛性，对模型训练过程的测试准确率进行评估。在响应时间分布为 LogNorm 和 Pareto 下，FedAsync 方法的准确率保持在较高水平，85.5%～90.1%。当 α 从 4 增加到 10 时，FedAlpha 方法的准确率受到越大的负面影响，逐渐从 89.3%下降到 83.4%。分析认为，FedAlpha 方法聚合所采用的全局模型不是最新的版本，存在迟滞效应，导致准确率下降。

总的来看，FedAlpha 方法可以在准确性略有下降的代价下，显著降低隔代模型泄露风险，提高异步联邦学习的安全性。

4 结论

本文提出一种异步联邦学习场景中隔代模型泄露攻击，并设计了一种基于 α-滑动平均的随机聚合算法。在隔代模型泄露攻击中，多名恶意客户端可相互串通，利用全局模型的版本差异，反推其它客户端的梯度信息。经研究发现，客户端的更新顺序是隔代模型泄露攻击的关键。基于 α-滑动平均的随机聚合算法能避免恶意客户端反推更新顺序，从而有效降低隔代模型泄露攻击的可能性。

本文的核心贡献如下：

1)首次提出模型隔代泄露攻击方式。在联邦学习训练过程中，当客户端按照特定更新顺序上传本地模型时，多个恶意客户端利用异步模型聚合算法，反向推算出其他客户端的本地模型。

2)针对模型隔代泄露攻击方式，提出基于 α-滑动平均的随机聚合算法，通过增加聚合过程的随机性，使得恶意客户端无法确定客户端更新顺序，从而确保恶意客户端无法反推正常客户端上传的本地模型。

3)仿真实验模拟 2 种不同响应时间分布下异步联邦学习框架的训练过程。经研究发现，随着恶意客户端的数量增加、全局迭代训练轮数增加时，恶意客户端发起隔代模型泄露攻击的次数将增大。基于 α-滑动平均的随机聚合算法可以有效降低异步联邦学习框架遭受隔代模型泄露攻击的可能性。

参考文献

[1] OQUAB M，BOTTOU L，LAPTEV I，et al. Flexible Clustered Federated Learning for Client-Level Data Distribution Shift[J]. IEEE Transactions on Parallel and Distributed Systems，2022，33(11)：2661-2674.

[2] 胡智尧，于淼，田开元. 面向半异步协同场景的联邦学习数据共享方法[J]. 海军工程大学学报，2024，36(3)：108-112.
HU Z Y，YU M，TIAN K Y. Data Sharing for Semi-Asynchronous Federated Learning[J]. Journal of Naval University of Engineering，2024，36(3)：108-112.

(in Chinese)

[3] ZHANG Z, ZHANG Y, GUO D, et al. Communication-Efficient Federated Continual Learning for Distributed Learning System with Non-IID Data[J]. Science China Information Sciences, 2022, 66(2): 122102.

[4] MCMAHAN H B, MOORE E, RAMAGED, et al. Communication-Efficient Learning of Deep Networks from Decentralized Data[C]//Proceedings of the International Conference on Artificial Intelligence and Statistics. Fort Lauderdale: PMLR, 2017: 1273-1282.

[5] ZHU L, LIU Z, HANS. Deep Leakage from Gradients [C]//Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2019: 14747-14756.

[6] PAILLIER P. Public-Key Cryptosystems Based on Composite Degree Residuosity Classes[M]//Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2007: 223-238.

[7] ZHENG S, MENG Q, WANGT, et al. Asynchronous Stochastic Gradient Descent with Delay Compensation [C]//Proceedings of the International Conference on Machine Learning. Sydney, NSW: PMLR, 2017: 4120-4129.

[8] ELOUADRHIRI A, ABDELHADI A. Differential Privacy for Deep and Federated Learning: a Survey[J]. IEEE Access, 2022, 10: 22359-22380.

[9] HENRY C, DAN B. Prio: Private, Robust, and Scalable Computation of Aggregate Statistics[C]//Proc of the 14th USENIX Symposium on Networked Systems Design and Implementation. Berkeley CA: USENIX, 2017: 259-282.

[10] CONG X, SANMI K, INDRANIL G. Asynchronous Federated Optimization[C]//NeurIPS Workshop on Optimization for Machine Learning. Cambridge: MIT-Press, 2020.

[11] ABADI M, BARHAM P, CHEN J, et al. Tensorflow: Large-Scale Machine Learning on Heterogeneous Distributed Systems [EB/OL]. (2016-03-16) [2022-11-04]. https://arxiv.org/abs/1603.04467.

(编辑：韩茜)

(上接第 36 页)

[14] FATURECHI R, LEVENBERG E, MILLER-HOOKS E. Evaluating and Optimizing Resilience of Airport Pavement Networks[J]. Computers & Operations Research, 2014, 43: 335-348.

[15] 民用机场水泥混凝土道面设计规范：MH/T 5004－2010[S]. 北京：中国民航出版社，2019.
Design Specification for Cement Concrete of Civil Airports: MH/T 5002－2010[S]. Beijing: Civil Aviation Press of China, 2019. (in Chinese)

[16] 建筑抗震韧性评价标准：GB/T 38591－2020[S]. 北京：中国标准出版社，2020.
Evaluation Criteria for Seismic Resilience of Buildings: GB/T 38591－2020[S]. Beijing: China Standard Press, 2020. (in Chinese)

[17] 安全韧性城市评价指南：GB/T 40947－2021[S]. 北京：中国标准出版社，2021.
Guidelines for Evaluating Safe and Resilient Cities: GB/T 40947－2021[S]. Beijing: China Standard Press, 2021. (in Chinese)

[18] 申晨. 城市公共交通网络空间拓扑结构特性及鲁棒性研究[D]. 哈尔滨：东北林业大学，2023.
SHEN C. Research on Topological Structure Characteristic and Robustness of Urban Public Transport Network Space[D]. Harbin: Northeast Forestry University, 2023. (in Chinese)

[19] 许钦钧，徐龙琴，刘双印. 一种基于网络拓扑的关键节点挖掘算法[J]. 仲恺农业工程学院学报，2023，36(2)：21-30.
XU Q J, XU L Q, LIU S Y. A Topology Based Heuristics Algorithm for Critical Node Detection[J]. Journal of Zhongkai University of Agriculture and Engineering, 2023, 36(2): 21-30. (in Chinese)

[20] 刘阳. 降雨内涝条件下城市道路网络韧性评价研究[D]. 哈尔滨：哈尔滨工业大学，2023.
LIU Y. Research on the Resilience Evaluation of Urban Road Network Under Rainfall and Water Logging Conditions[D]. Harbin: Harbin Institute of Technology, 2023. (in Chinese)

(编辑：刘勇)