软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2024,35(7):3454−3468 [doi: 10.13328/j.cnki.jos.006914]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于模型后门的联邦学习水印*

李 璇[1,2,3], 邓天鹏[1], 熊金波[1,2], 金 彪[1], 林 劼[1]

[1](福建师范大学 计算机与网络空间安全学院, 福建 福州 350117)

[2](福建省网络安全与密码技术重点实验室, 福建 福州 350117)

[3](数字福建大数据安全技术研究所, 福建 福州 350117)

通信作者: 李璇, E-mail: jessieli24@fjnu.edu.cn

摘 要: 高精度联邦学习模型的训练需要消耗大量的用户本地资源, 参与训练的用户能够通过私自出售联合训练的模型获得非法收益. 为实现联邦学习模型的产权保护, 利用深度学习后门技术不影响主任务精度而仅对少量触发集样本造成误分类的特征, 构建一种基于模型后门的联邦学习水印 (federated learning watermark based on backdoor, FLWB) 方案, 能够允许各参与训练的用户在其本地模型中分别嵌入私有水印, 再通过云端的模型聚合操作将私有后门水印映射到全局模型作为联邦学习的全局水印. 之后提出分步训练方法增强各私有后门水印在全局模型的表达效果, 使得 FLWB 方案能够在不影响全局模型精度的前提下容纳各参与用户的私有水印. 理论分析证明了 FLWB 方案的安全性, 实验验证分步训练方法能够让全局模型在仅造成 1% 主任务精度损失的情况下有效容纳参与训练用户的私有水印. 最后, 采用模型压缩攻击和模型微调攻击对 FLWB 方案进行攻击测试, 其结果表明 FLWB 方案在模型压缩到 30% 时仍能保留 80% 以上的水印, 在 4 种不同的微调攻击下能保留 90% 以上的水印, 具有很好的鲁棒性.

关键词: 联邦学习; 产权保护; 模型水印; 后门任务; 模型聚合

中图法分类号: TP309

中文引用格式: 李璇, 邓天鹏, 熊金波, 金彪, 林劼. 基于模型后门的联邦学习水印. 软件学报, 2024, 35(7): 3454–3468. http://www.jos.org.cn/1000-9825/6914.htm

英文引用格式: Li X, Deng TP, Xiong JB, Jin B, Lin J. Federated Learning Watermark Based on Model Backdoor. Ruan Jian Xue Bao/Journal of Software, 2024, 35(7): 3454–3468 (in Chinese). http://www.jos.org.cn/1000-9825/6914.htm

Federated Learning Watermark Based on Model Backdoor

LI Xuan[1,2,3], DENG Tian-Peng[1], XIONG Jin-Bo[1,2], JIN Biao[1], LIN Jie[1]

[1](School of Computer and Cyber Security, Fujian Normal University, Fuzhou 350117, China)

[2](Fujian Provincial Key Lab of Network Security & Cryptology, Fuzhou 350117, China)

[3](Digital Fujian Institute of Big Data Security Technology, Fuzhou 350117, China)

Abstract: The training of high-precision federated learning models consumes a large number of users' local resources. The users who participate in the training can gain illegal profits by selling the jointly trained model without others' permission. In order to protect the property rights of federated learning models, this study proposes a federated learning watermark based on backdoor (FLWB) by using the feature that deep learning backdoor technology maintains the accuracy of main tasks and only causes misclassification in a small number of trigger set samples. FLWB allows users who participate in the training to embed their own private watermarks in the local model and then map the private backdoor watermarks to the global model through the model aggregation in the cloud as the global watermark for federated learning. Then a stepwise training method is designed to enhance the expression effect of private backdoor watermarks in the

* 基金项目: 国家自然科学基金 (62272103, 62272102, 61872090)
收稿时间: 2021-12-05; 修改时间: 2022-11-03; 采用时间: 2023-02-06; jos 在线出版时间: 2023-08-23
CNKI 网络首发时间: 2023-08-28

global model so that FLWB can accommodate the private watermarks of the users without affecting the accuracy of the global model. Theoretical analysis proves the security of FLWB, and experiments verify that the global model can effectively accommodate the private watermarks of the users who participate in the training by only causing an accuracy loss of 1% of the main tasks through the stepwise training method. Finally, FLWB is tested by model compression and fine-tuning attacks. The results show that more than 80% of the watermarks can be retained when the model is compressed to 30% by FLWB, and more than 90% of the watermarks can be retained under four different fine-tuning attacks, which indicates the excellent robustness of FLWB.

Key words: federated learning (FL); property rights protection; model watermark; backdoor task; model aggregation

据 IDC 统计, 尽管受 COVID-19 的影响, 2020 年内全球依然产生了约 64.2 ZB 的数据. 这些海量数据赋能深度学习 (deep learning, DL) 模型将极大地提升智能设备的智能化程度[1]. 然而, 由于个人数据的高度敏感性, 将用户的本地数据毫无保留地发送到中心化服务器训练 DL 模型会增加隐私泄露的风险[2]. 为了应对这种挑战, 联邦学习 (federated learning, FL) 提供了一种分布式学习框架, 采用终端用户本地训练模型, 并将模型参数上传到第三方服务器进行聚合的方式, 生成有效的全局模型, 避免了用户数据的直接交互, 从而保护用户的数据隐私[3].

联邦学习按照训练方式的不同, 通常分为横向联邦学习、纵向联邦学习和联邦迁移学习[1]. 尽管训练方式不同, 考虑到数据的私有化, 训练时均需将模型下发到用户端进行本地训练, 再将本地训练的该轮模型权重上传至服务器进行聚合. 参与训练的本地用户通常使用笔记本、手机等便携设备进行训练, 这将会消耗大量的计算资源而导致较高的训练开销[4]. 同时, 由于每轮聚合后的全局模型无法保证达到优于用户本地模型的效果, 普通用户在综合考虑算力、通信开销等资源消耗与模型效果之后, 会降低参与训练的积极性, 甚至拒绝参与下一轮训练. 为激励用户参与联邦学习系统, 已有方案提出基于拍卖的方法为参与训练的用户给予奖励[5], 并通过历史的上传信息评估高质量用户参与聚合. 文献 [6] 提出了一种分布式算法, 使用户能够在不知道彼此对模型的估值和成本的情况下获得最大化收益. 但这些方案仅针对 FL 系统和用户群体的付出与收益是否对等的情况进行研究. 当参与中心服务器 A 主持的联邦学习训练中出现自私用户 m 时, m 能够在未经其他用户同意的情况下将训练好的模型 W_G 私自出售给中心服务器 B, B 还可以进行二次销售, 在获利的同时导致模型被滥用. 该过程侵犯了参与训练用户的知识产权, 会破坏用户参与联邦学习系统的积极性. 若用户无法通过有效的方式证明 B 所持有的模型为自己参与训练的模型 W_G, 则难以追责. 除此之外, 被恶意出售的模型能够被任意地拷贝、转发, 甚至能够通过已训练好模型的输出结果进行逆向攻击来恢复原始训练数据[7], 从而泄露用户隐私.

联邦学习环境下参与用户的数量较多, 难以避免上述存在自私用户恶意出售模型的情况. 因此, 设计有效的联邦学习模型知识产权保护方法, 通过验证各用户对于联邦学习模型的所有权, 以支持恶意出售模型情况发生时的诚实用户维权操作, 保障诚实用户的权益, 能够大大提高用户参与联邦学习训练的积极性. 数字水印技术作为常用的知识产权保护手段, 能够将认证信息嵌入到载体信息中, 以证明用户对于该载体信息的所有权[8]. 目前研究的常用信息载体为多媒体数据, 包括图像[9]、声音[10]、视频[11]等. 深度学习模型作为新型信息载体, 通过对样本数据进行训练, 将样本信息转化为模型参数, 同样需要进行知识产权保护[12]. 如何为深度学习模型嵌入水印, 防止模型被自私用户恶意泄露, 仍处于起步阶段. 文献 [13] 针对深度学习分类问题, 使用黑盒的方式利用标签翻转对深度学习模型嵌入水印, 并通过理论证明和实验验证方案的安全性. 文献 [14] 通过对抗学习的方式向白盒模型嵌入水印, 在不影响模型精度的前提下进一步提高了水印的隐蔽性和鲁棒性. 文献 [15] 针对模型窃取攻击提出 DAWN 算法来进行水印嵌入, DAWN 不修改模型训练过程, 仅在用户调用查询 API 时将响应结果集合的子集作为水印, 当恶意用户利用该结果集合训练替代模型时将水印嵌入到替代模型中. 文献 [16] 针对模型水印可能遇到的模糊攻击问题, 提出基于 passport 的水印算法, 根据用户输入的 passport 的真伪来改变已发布模型的推理阶段的性能, 从而让模型在遇到伪造 passport 时性能变差. 文献 [17] 则对生成对抗网络的产权保护问题, 提出 2 种完整的保护算法: 在黑盒模式下构建重构正则化算法, 以允许生成器在给定触发输入时在合成图像的指定位置嵌入水印; 在白盒模式下对文献 [16] 的符号损失进行改进, 使得能够根据归一化层的缩放因子 γ 的符号提取出有效信息. 上述算法均是针对深度学习模型进行单水印的嵌入. 在联邦学习环境中, 虽然其底层采用的是深度学习模型, 但由于参与联邦学习训练的用户存在多个, 若每个用户直接应用传统深度学习水印算法于本地训练结果, 则在服务器端聚合

模型时会影响各用户水印的表现, 导致无法有效地将多用户水印成功嵌入联邦学习全局模型中. 横向联邦学习 (horizontal federated learning, HFL) 是目前研究最为广泛的 FL 系统, 其各方用户获得同样的全局模型, 参与训练用户量多, 且用户可信性难以衡量, 迫切需要对模型进行产权保护. 针对 HFL 的产权保护问题, 文献 [18] 结合基于特征和基于后门的两种水印嵌入方式提出 FedIPR 算法, 后门水印嵌入时要采用额外的神经网络使用 PGD 方案生成对抗样本, 再用其训练后门水印模型, 需要参与训练用户具备相应知识, 同时增加了联邦学习客户端的操作复杂度. 本文针对横向联邦学习场景, 构建了一种基于模型后门的联邦学习水印方案 (federated learning watermark based on backdoor, FLWB), 并从理论和实验上验证了该方案的安全性和有效性. 主要贡献如下.

(1) 设计一种基于模型后门的联邦学习水印方案. 通过用户本地训练私有后门模型, 让各本地模型在正常数据的训练任务 (正常任务) 上保持高精度而在后门触发集数据的训练任务 (后门任务) 上产生误分类, 使得服务器端在联邦学习的聚合阶段将各用户的私有后门水印融入全局模型, 实现私有后门到联邦学习全局水印的映射. 该方案中后门样本的筛选是在水印模型整体训练时随机执行标签翻转得到, 无需使用额外的神经网络训练后门样本, 更加便捷且随机性高.

(2) 设计分步训练方法, 缓解各用户本地模型间水印选择的不同而存在的冲突问题. 经过多轮迭代, 参与联邦学习训练用户的后门水印可以在全局模型中得以保留, 以支持后续的模型知识产权验证. 分步训练方法使得方案对于后门样本的训练更加可控, 能够提高指定水印嵌入模型的成功率. 同时, 采用承诺方案对水印验证的不可伪造性进行增强, 提供了理论保证.

(3) 对 FLWB 方案抵御攻击的安全性进行了形式化描述和理论证明. 实验验证 FLWB 方案和分步训练方法的有效性, 并使用常见的模型压缩方法和模型微调方法作为攻击手段, 在不影响全局模型准确性的前提下对水印进行攻击, 验证 FLWB 方案的鲁棒性.

本文第 1 节对 FLWB 中所涉及的符号和概念进行定义, 并对承诺方案进行介绍. 第 2 节介绍 FLWB 方案的构建方法和算法步骤, 同时对其性质进行定义. 第 3 节对 FLWB 方案进行安全性分析和性能评估, 证明 FLWB 方案的安全性、有效性和鲁棒性. 第 4 节总结全文并对未来研究方向进行展望.

1 模型定义与问题描述

本节首先对贯穿全文的符号进行说明, 之后分别为联邦学习和联邦学习后门进行形式化定义, 在本节的最后简述需要的预备知识.

1.1 符号说明

FLWB 方案所用符号的汇总与说明如表 1 所示.

表 1 符号汇总与说明

符号	描述	符号	描述
$p\in\mathbb{N}$	p 为系统安全参数, 是所有算法的隐式输入	M_i	各用户本地训练所得模型
PPT	概率多项式时间算法	M	本地模型聚合后全局模型
$f(\cdot)$	分类准确率为100%的理想化分类函数	T	触发集数据
O_f	能对询问进行真实回复的随机预言机	T_L	后门标签
N	用户数量, 各用户编号分别为 $1,2,\ldots,N$	$\mathcal{B}$	后门数据集 $\mathcal{B}=(T,T_L)$
D_i	用户 i 持有的数据集	$\hat{M}$	被后门任务所标记的全局后门模型
L	数据集的标签集合	(mk_i,vk_i)	mk_i 为用户 i 的水印标记密钥, vk_i 为其验证密钥
$\perp$	空集, 代表未定义的标签名	(MK,VK)	N 个用户密钥 (mk_i,vk_i) ($i=1,\ldots,N$) 的集合

1.2 联邦学习语义描述

假设存在某个客观真实的函数 f, 它根据一个固定的输出标签集来对输入进行分类, 倘若该标签为未定义, 则

记为 $\perp$. 联邦学习作为一种隐私保护的分布式协作学习方法, 允许多个用户共同训练以达成目标任务. 本文主要针对深度学习中的分类问题, 将联邦学习任务分为 2 个步骤[2]: 联合训练过程 (federated training, FedTrain) 和推理预测过程 (inferential prediction, InferPre). FedTrain 是通过各用户本地训练 (train) 私有数据集, 使得聚合模型所拟合的函数 f' 具有与真实分类函数 f 相似的能力, 并且允许 InferPre 能够在未训练过的数据上表现良好 (分类正确). 下面对深度学习的语义描述[13]进行扩展, 以适应联邦学习环境, 如图 1 所示.

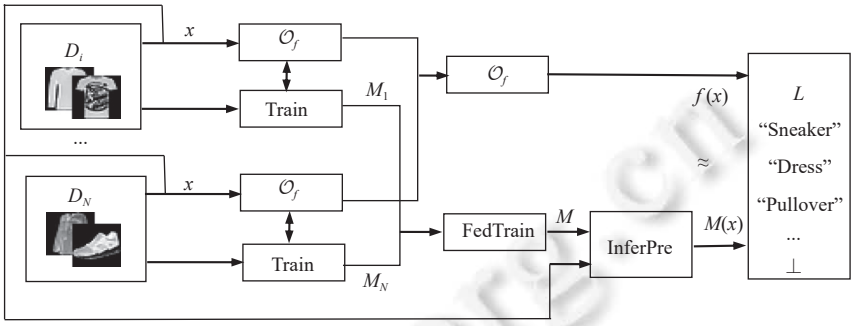


图 1 联邦学习 FedTrain 语义描述图

图 1 描述的联邦学习系统中, 各参与用户的数据集是本地私人持有. 设用户 $i \in \{1,2,\ldots,N\}$ 持有转换成 0/1 编码的数据集 $D_i \subset \{0,1\}^*$, 对应标签集为 $L \subset \{0,1\}^* \cup \{\perp\}$. $D = D_1 \cup \ldots \cup D_N$ 是所有可能的输入样本组成的输入集合, L 是给定输入所对应的标签集合且为其所代表标签值的 one-hot 编码. 同时, $\perp \in L$ 表示当前样本未被分配给一个指定输出即暂无标签的样本.

假设存在一个理想的函数 f 能够将指定标签分配给输入数据, 即存在理想状态的分类器 $f: D \to L$ 使得不同样本能够被完美识别到不同的类别. 当对于特定的任务和标签时, f 可能是未定义的, 故本文将有真实标签的数据集记为 $\tilde{D} = \{x \in D | f(x) \neq \perp\}$, 且 $\tilde{D} = \tilde{D}_1 \cup \ldots \cup \tilde{D}_N$. 为形式化定义 FL 学习过程, 规定学习算法仅能通过随机预言机 $\mathcal{O}_f$ 访问 f, 并且随机预言机 $\mathcal{O}_f$ 将诚实地回答关于 f 的询问. 在 FL 设置中, 各用户拥有相同的学习目标 f, 同时本文假设用户数据分布为独立同分布, 则各方所能获取的随机预言机 $\mathcal{O}_f$ 相同.

由此, 完整的 FL 任务可以表示为 FedTrain 和 InferPre 两个步骤.

(1) FedTrain($\mathcal{O}_f$) 是一个概率多项式时间算法, 能够让各用户在 $p(n)$ 时间内本地训练其私有数据集 (train), 并在云端聚合器聚合 (aggregator) 后输出模型 $M \subset \{0,1\}^{p(n)}$.

(2) InferPre(M, x) 是一个确定性多项式时间算法, 能够对于给定输入 $x \in D$ 输出 $M(x) \in L \backslash \{\perp\}$.

故对于给定函数 f, 若 FL 训练结束后全局模型满足 $\Pr_{x \in \tilde{D} \backslash T}[f(x) \neq \text{InferPre}(M, x)] \leqslant \epsilon$, 则称联邦学习算法 (FedTrain, InferPre) 满足 $(1-\epsilon)$-精度.

1.3 联邦学习后门描述

本文将通过后门技术实现 FL 模型中多用户水印的嵌入, 本节先引入联邦学习后门的形式化描述. 联邦学习 FL 作为一种隐私保护的分布式学习系统, 其底层学习算法依赖于不断发展的机器学习算法. 后门神经网络 (backdooring neural network, BNN) 是一种训练机器学习模型使得故意将原始标签输出为指定错误目标标签的技术[19]. 假设将给定输入子集 $T \subset D$ 定义为触发集, BNN 能够通过函数 $T_L: T \to L \backslash \{\perp\}; x \mapsto T_L \neq f(x)$ 捕捉基于真实 f 的错误标签, 函数 T_L 不允许输出空标签 $\perp$. 触发集 T 与标记函数 T_L 共同组成后门 $\mathcal{B} = (T, T_L)$, 触发集与标记函数往往成对出现, 故在后文指定一个触发集 T 时, 标记函数 T_L 将被隐式地定义.

对于一个后门 $\mathcal{B}$, 后门算法 Backdoor($\mathcal{O}_f, \mathcal{B}, M$) 是一个概率多项式时间 PPT 算法, 能够在模型输入时输出一个在触发集上有高概率错误分类的模型 $\hat{M}$. 若 $\hat{M}$ 能够对触发集 T 中以高概率输出指定标签, 而在正常样本集 $\tilde{D} \backslash T$ 中也表现良好, 则 $\hat{M}$ 称为被后门任务所标记, 即:

$$\begin{cases} \Pr\limits_{x\in\tilde{D}\backslash T}\left[f(x)\neq \mathrm{InferPre}\left(\hat{M},x\right)\right]\leqslant\epsilon \\ \Pr\limits_{x\in T}\left[T_L(x)\neq \mathrm{InferPre}\left(\hat{M},x\right)\right]\leqslant\epsilon \end{cases} \tag{1}$$

在联邦学习中, 各用户作为模型训练的个体会独立生成后门触发集, 各用户希望所生成的本地后门模型在全局聚合后在触发集上保留高概率错误分类, 即用户 $1,2,\ldots,N$ 分别持有后门 $\mathcal{B}_1=(T_1,T_{1,L})$, $\mathcal{B}_2=(T_2,T_{2,L}),\ldots,\mathcal{B}_N=(T_N,T_{N,L})$, 则最终生成的全局后门模型 $\hat{M}$ 能够在 $T=T_1\cup T_2\cup\ldots\cup T_N$ 中有公式 (1) 效果. 本文方案针对横向联邦学习场景进行设计, 由于横向联邦学习用户数据集的私有性, 在此假设各用户所选触发集相交的概率可忽略不计, 即对于任意用户 i、j $(i\neq j)$ 的后门触发集, 满足 $\Pr[\mathcal{B}_i\cap\mathcal{B}_j\neq\varnothing]<\epsilon$. 为综合考虑触发集的选择与联邦学习任务, 本文将触发集的生成算法 SampleBackdoor 并入联邦学习算法 (FedTrain, InferPre) 中. 图 2 是联邦学习后门训练的语义示意图.

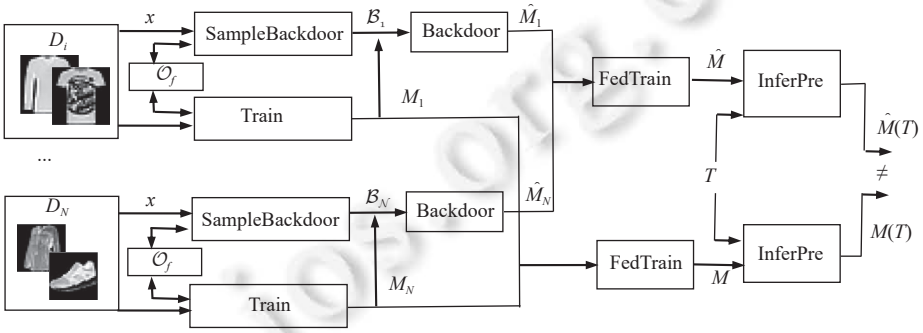


图 2 联邦学习后门训练语义示意图

后门任务利用深度学习模型参数的冗余性, 通过训练过程将对正确分类起到正面影响的部分参数和不起作用的冗余参数变成负面影响以植入后门任务, 而其他样本依然能够正确分类. 本文利用后门的这个特性作为水印嵌入的依据, 即仅在选定的水印图像上输出指定的错误标签以与正常输出的样本区分, 由此构建后门任务与水印的映射.

后门 (水印) 触发集的选择在具有随机性的同时应该具有鲁棒性, 即攻击者 $\mathcal{A}$ 在不具备关于触发集 T 的先验知识的情况下难以去除该后门 (水印). 在获取到带有后门 (水印) 的联邦学习系统后, $\mathcal{A}$ 会试图移出后门 (水印) 以声明关于该模型的所有权, 为此其可以采取攻击的形式破坏后门 (水印) 任务的完整性. 本文对攻击者 $\mathcal{A}$ 的能力进行以下刻画.

(1) $\mathcal{A}$ 不具备无限的算力和不限次数的访问 O_f 的能力, 其操作仅能在 PPT 时间内进行.

(2) $\mathcal{A}$ 所伪造的模型需保证分类任务的精度, 其生成的不含后门任务的伪造模型 M' 应与 $\hat{M}$ 不可区分, 即 $\Pr\left[M'(x)]\neq\hat{M}(x)\right]<\epsilon$.

1.4 承诺方案

承诺方案是一种广泛使用的密码学原语[20], 允许发送方 (sender, S) 将一个秘密 x 锁定到一个无泄露、防篡改的加密库中, 并将其交给接收者 (receiver, R). 承诺方案有 2 个基本性质: 隐藏 (hiding) 和绑定 (binding). R 无法在没有 S 的帮助下打开这个秘密 (隐藏), S 也无法通过其他方式将消息泄露给其他人或更改被隐藏的秘密 x (绑定).

一个承诺方案由 (*Com*, *Open*) 这 2 个算法组成, 其正式定义如下.

(1) $Com(x,r)$: 给定输入 $x\in S$ 和随机比特串 $r\in\{0,1\}^n$, 输出比特串 c_x, 其中 n 为随机比特串长度.

(2) $Open(c_x,x,r)$): 给定 $x\in S$, $r\in\{0,1\}^n$, $c_x\in\{0,1\}^*$, 输出 0 或 1.

承诺方案的相应性质如下.

(1) 正确性: 对于给定承诺 $c_x\leftarrow Com(x,r)$, 在相应比特串 r 下打开该承诺的概率为 1, 即:

$$\Pr_{r\in\{0,1\}^n}[Open(c_x,x,r)=1]=1,\forall x\in S \tag{2}$$

(2) 绑定: 对于任一 PPT 时间算法 Adv 生成的 $(\tilde{x},\tilde{r})$ 且 $(\tilde{x},\tilde{r})\neq(x,r)$, 无法打开承诺 $c_x\leftarrow Com(x,r)$, 即:

$$\Pr[Open(c_x,\tilde{x},\tilde{r})=1]\leqslant\epsilon(n) \tag{3}$$

其中, $\epsilon(n)$ 对于 n 可忽略不计, 则称 (*Com*, *Open*) 是绑定的.

(3) 隐藏: 若对于 $\forall x\in S$, $r\in\{0,1\}^n$, 不存在 PPT 时间算法 Adv 能够区分 $c_0\leftarrow Com(0,r)$ 和 $c_x\leftarrow Com(x,r)$, 则称 (*Com*, *Open*) 是隐藏的. 当 c_0 与 c_x 的统计分布相近时, 则称该承诺方案是统计上隐藏[21].

2 联邦学习水印 (FLWB) 方案构造

本节首先给出联邦学习水印的形式化定义, 再对联邦学习水印方案需要满足的基本性质进行描述, 最后提出基于模型后门的联邦学习水印方案.

2.1 FLWB 定义

联邦学习模型的训练数据以私有化的形式存储在本地, 为保证 FLWB 中水印的私有化, 本文将文献 [15] 定义的深度学习水印扩展到联邦学习环境, 并将 FLWB 方案拆分成 3 个算法, 分别称为联邦密钥生成算法 FedKeyGen、联邦标记算法 FedMark 和验证算法 Verify. 其形式化定义如下.

(1) FedKeyGen(p): 给定安全参数 p, 用户 i 输出密钥对 (mk_i, vk_i) 且保证密钥对私有存储, 记所有用户的密钥对集合为 $(MK,VK)=\cup_{i=1}^{N}(mk_i,vk_i)$.

(2) FedMark(M,mk_i): 用户 i 对于给定输入模型 M 和标记密钥 mk_i, 输出后门水印模型 $\hat{M}_i$, 记聚合后的全局后门水印模型为 $\hat{M}$.

(3) Verify(MK, VK, M'): 对输入密钥对 (MK, VK) 中的每一组密钥 (mk_i, vk_i) 进行验证, 每一组输出 1 比特信息 $b_i\in\{0,1\}$, 0 表示该模型不含水印, 1 表示含有水印. 最终验证输出 1 比特信息 $b=b_1\wedge b_2\wedge\ldots\wedge b_N$.

上述水印方案涉及的 3 种算法均是 PPT 时间算法. 其中, (MK,VK) 表示所有用户的标记-验证密钥对的集合, 每个用户的密钥对 (mk_i,vk_i) 仅本地存储并未广播给其他用户, 即每个用户仅知道 (MK,VK) 中属于自己的密钥对. 本文所提 FLWB 方案整体嵌入流程如算法 1 所示.

算法 1. 联邦学习水印算法 MarkModel().

输入: O_f;
输出: MK, VK, M, $\hat{M}$.

1. 生成模型 $M\leftarrow\text{FedTrain}(O_f)$.
2. 用户 i 生成密钥对 $(mk_i,vk_i)\leftarrow\text{FedKeyGen}()$.
3. 用户 i 本地训练模型 $\hat{M}_i\leftarrow\text{FedMark}(M,mk_i)$.
4. 输出 $\left(M,\hat{M},MK,VK\right)$.

算法 1 是 FLWB 方案中嵌入算法的理论构建, 表明在对给定分类问题上能够通过 FedTrain 算法训练出 $(1-\epsilon)$ 精度模型 M, 且负责密钥生成的 FedKeyGen 和负责水印标记的 FedMark 能够为 M 嵌入水印得到 $\hat{M}$.

2.2 FLWB 性质

考虑到联邦学习的底层模型使用了深度学习算法, FLWB 方案应该维持模型功能性, 并且同时具备传统水印方案的正确性、不可移除性、不可伪造性及归属权的平凡性. 以下将对所提性质进行形式化描述. 其中, 攻击者所进行的攻击算法 $\mathcal{A}$ 仅能够在其本地数据集中进行操作.

性质 1. 正确性. 算法 (FedKeyGen, FedMark, Verify) 应该保证诚实用户的标记密钥能够通过验证, 即:

$$\Pr_{(M,\hat{M},MK,VK)\leftarrow \text{MarkModel}()}\left[\text{Verify}\left(MK,VK,\hat{M}\right)=1\right]=1 \tag{4}$$

性质 2. 功能保持性. 最终模型表现能力在有水印的情况下和无水印模型效果相似, 即对于 $\forall\left(M,M,\hat{M},MK,VK\right)\leftarrow \text{MarkModel}()$, 公式 (5) 成立.

$$\left|\Pr_{x\in\tilde{D}}\left[\text{InferPre}\left(x,M\right)=f\left(x\right)\right]-\Pr_{x\in\tilde{D}}\left[\text{InferPre}\left(x,\hat{M}\right)=f\left(x\right)\right]\right|<\epsilon \tag{5}$$

其中, ϵ 为可忽略小数.

性质 3. 不可移除性. 攻击者即使在知道水印存在, 甚至了解所用水印嵌入算法的情况下仍无法移除所嵌入的水印信息. 这要求对于任何 PPT 时间算法 $\mathcal{A}$, 敌手赢得下述游戏的优势是可忽略的.

i. 计算 $\left(M,\hat{M},MK,VK\right)\leftarrow \text{MarkModel}()$.

ii. 执行 $\mathcal{A}$ 并计算 $\tilde{M}\leftarrow\mathcal{A}\left(O_f,\hat{M},VK\right)$.

iii. $\tilde{M}$ 的精度满足 $\Pr_{x\in\tilde{D}}\left[\text{InferPre}\left(x,M\right)=f\left(x\right)\right]\approx\Pr_{x\in\tilde{D}}\left[\text{InferPre}\left(x,\tilde{M}\right)=f\left(x\right)\right]$ 且 $\text{Verify}\left(MK,VK,\tilde{M}\right)=0$.

性质 4. 不可伪造性. 当攻击者知道验证密钥 VK 而不知道标记密钥 MK 时, 无法向第三方证明攻击者对于该模型的所有权. 即对于任意 PPT 时间算法 $\mathcal{A}$, 敌手赢得下述游戏的优势是可忽略的.

i. 计算 $\left(M,\hat{M},MK,VK\right)\leftarrow \text{MarkModel}()$.

ii. 敌手执行 $\left(\tilde{M},\tilde{MK}\right)\leftarrow\mathcal{A}\left(O_f,\hat{M},VK\right)$.

iii. $\text{Verify}\left(\tilde{MK},VK,\tilde{M}\right)=1$.

性质 5. 非平凡所有权. 任一攻击者即使在知道水印的嵌入算法时, 不能预先生成一个密钥对 (MK,VK), 使其能够声称其所不知道的任意模型的所有权. 即对于任意 PPT 算法 $\mathcal{A}$, 敌手赢得下述游戏的优势是可忽略的.

i. $\mathcal{A}$ 伪造标记密钥与验证密钥对 $\left(\tilde{MK},\tilde{VK}\right)$.

ii. 计算 $\left(M,\hat{M},MK,VK\right)\leftarrow \text{MarkModel}()$.

iii. $\text{Verify}\left(\tilde{MK},\tilde{VK},\hat{M}\right)=1$.

2.3 FLWB 方案构建

基于第 1.3 节联邦学习后门任务的介绍和第 2.1 节联邦学习水印的定义, 本节将利用深度学习后门任务实现 FLWB 方案的构建. 同时使用承诺方案增强 FLWB 整体方案的完整性和可验证性. 本文的目标是在 FL 系统训练结束时输出一个带有所有参与联邦训练的用户水印的全局模型 $\hat{M}$.

FLWB 方案将利用深度学习后门任务仅在少量触发集样本产生误分类的特征, 通过联邦学习的聚合阶段将各用户的私有后门融入全局模型, 从而形成私有后门任务到联邦学习全局模型水印的映射, 达到保护模型所有权归属的目的. 在底层上, 由于正常的深度学习模型在训练结束后, 不同的神经元会对不同输入数据的特征产生激活或抑制状态, 从而影响该样本的分类结果. 后门任务则利用深度学习模型参数的冗余性, 将本应对后门触发集数据输入时产生抑制状态的神经元变为激活状态, 改变了对该样本的决策结果, 使之进行错误输出. 由于联邦学习中的每个用户仅能训练本地模型, 每次迭代聚合过程中每个本地模型仅会对全局模型产生轻微的影响, 甚至不同本地模型在本次迭代聚合后的效果会抵消, 故只有经过多轮迭代聚合才能对全局模型造成较为稳定的后门影响[22]. 所以, 从整体性上看 FLWB 是在全局模型中以嵌入后门任务的形式注入水印, 但在较低层次上, 该水印的嵌入需要通过各用户在本地模型上执行后门任务后, 通过云端聚合使得各用户后门映射到全局模型. 因此, 各用户所构建的后门任务本身即为标记密钥 mk_i, 而对后门任务所用触发集进行承诺方案绑定的值即为验证密钥 vk_i.

下面对 FLWB 方案进行详细描述. 假设联邦学习算法 (FedTrain, InferPre) 是 $(1-\epsilon)$ 精度的模型, Backdoor 是后门算法, 且 (*Com*, *Open*) 为统计上隐藏的承诺方案, 则 FLWB 方案的 3 个子算法 (FedKeyGen, FedMark, Verify) 如算法 2、算法 3 和算法 4 所示.

算法 2. 联邦学习水印密钥生成算法 FedKeyGen().

输入: 安全参数 p, 随机比特串长度 n;
输出: MK, VK.

1.用户 i 执行 $\mathcal{B}_i=(T_i,T_{i,L})\leftarrow \text{SampleBackdoor}(O_f)$,其中 $T_i=\left\{t_i^{(1)},\ldots,t_i^{(n)}\right\}$, $T_{i,L}=\left\{T_{i,L}^{(1)},\ldots,T_{i,L}^{(n)}\right\}$.记 $(T,T_L)=\cup_{i=1}^{N}(T_i,T_{i,L})$.
2. 用户 i 随机生成长为 $2n$ 的比特串 $r_{i,t}^{(j)},r_{i,L}^{(j)}\leftarrow\{0,1\}^n$, 并生成承诺 $\left\{c_{i,t}^{(j)},c_{i,L}^{(j)}\right\}_{j\in[n]}$, 其中 $c_{i,t}^{(j)}\leftarrow \text{Com}\left(t_i^{(j)},r_{i,t}^{(j)}\right)$, $c_{i,L}^{(j)}\leftarrow \text{Com}\left(T_{i,L}^{(j)},r_{i,L}^{(j)}\right)$.
3. 记用户 i 的标记密钥 mk_i 和验证密钥 vk_i 分别为: $mk_i\leftarrow\left(\mathcal{B}_i,\left\{r_{i,t}^{(j)},r_{i,L}^{(j)}\right\}_{j\in[n]}\right)$, $vk_i\leftarrow\left\{c_{i,t}^{(j)},c_{i,L}^{(j)}\right\}_{j\in[n]}$. 记 $(MK,VK)\leftarrow\cup_{i=1}^{N}(mk_i,vk_i)$.

算法 3. 联邦学习水印标记算法 FedMark().

输入: M, MK;
输出: $\hat{M}$.

1. 用户 i 收到模型 M, 记 $mk_i=\left(\mathcal{B}_i,\left\{r_{i,t}^{(j)},r_{i,L}^{(j)}\right\}_{j\in[n]}\right)$.
2. 用户 i 在第 t 轮训练模型并输出 $\hat{M}_i^t\leftarrow \text{Backdoor}(O_f,\mathcal{B}_i,M)$.
3. 云端输出聚合模型 $\hat{M}^t=\text{Aggregator}\left(\hat{M}_1^t,\ldots,\hat{M}_N^t\right)$.
4. 重复步骤 2 和步骤 3 直到完成规定训练轮数, 则最终模型为 $\hat{M}$.

算法 4. 联邦学习水印验证算法 Verify().

输入: MK, VK, M;
输出: 0/1.

1. 对于 $\forall mk\in MK$, $\forall vk\in VK$, $(b,b_L)\in(T,T_L)$, 验证 $\forall t^{(j)}\in b: b_L^{(j)}\neq f(t^{(j)})$ 是否成立. 若成立则输出 1, 否则输出 0.
2. 对于 $\forall i\subset N$, $\forall j\in\{1,2,\ldots,p\}$, 验证 $Open\left(c_{i,t}^{(j)},t_i^{(j)},r_i^{(j)}\right)=1$ 和 $Open\left(c_{i,L}^{(j)},T_{i,L}^{(j)},r_{i,L}^{(j)}\right)=1$ 是否成立. 若成立则输出 1, 否则为 0.
3. 对于 $\forall i\subset N$, $\forall j\in\{1,2,\ldots,p\}$, 验证 $\text{InferPre}\left(t_i^j,\hat{M}\right)=T_{i,L}^{(j)}$ 是否成立. 若上式对 T_i 中除 $\epsilon|T_i|$ 个元素外成立, 则输出 1, 否则输出 0.

上述 3 个子算法完整地构建了 FLWB 方案, 各用户通过随机选择本地后门触发集, 在接收到当前轮全局模型 $M^{(t)}$ 后训练本地后门模型 $\hat{M}_i^{(t)}$, 通过云端聚合将后门任务映射到全局模型 $\hat{M}^{(t+1)}$.

后门任务的训练过程主要体现在 FedMark 算法的 Backdoor 中, 本文在用户本地训练阶段采用分步训练方法将正常样本的训练过程与触发集样本训练解耦, 通过在每一批次的正常样本训练之后单独训练触发集样本, 以提高触发集样本的成功率. Backdoor 算法的具体步骤详见算法 5.

算法 5. 后门水印嵌入算法 Backdoor().

输入: 用户 i 的后门数据集 $\mathcal{B}_i$, 待训练模型 M;
输出: 用户 i 的含水印模型 $\hat{M}_i$.

1. 用户 i 使用前一轮全局模型 M_G^{t-1} 作为当前训练的初始模型.
2. 使用交叉熵损失函数训练正常样本集 $D_i\backslash\mathcal{B}_i$, 记其中第 r 轮本地训练的模型参数为 w_r.
3. 在 w_r 的基础上使用交叉熵损失函数训练触发集样本 $\mathcal{B}_i$, 记模型参数为 $\tilde{w}_r$.

4. 计算后门样本的模型更新值 $\delta_r = \tilde{w}_r - w_r$.

5. 将 δ_r^t 增强 λ 倍后加入 w_r 中, 即 $M_r^t = w_r + \lambda\delta_r$.

6. 将 M_r^t 作为用户 i 第 $r+1$ 轮本地训练的初始模型, 重复步骤 2–5 直到模型在 $\mathcal{B}_i$ 的损失为 0, 则用户 i 在该轮全局训练的模型为 $\hat{M}_i$.

由于联邦学习训练是个动态的交互过程, 在云端与用户的不断迭代通信中得到最终模型, 使得 FL 系统在训练结束时输出一个带有所有用户私有后门任务的全局模型 $\hat{M}$. 本文在 FedMark 算法中使用的聚合器 Aggregator 为 Federated Averaging[2]. 后文将对 FLWB 方案的安全性进行证明, 并验证该方案的有效性.

3 FLWB 方案安全性分析与性能评估

本节将对 FLWB 方案的安全性、有效性和鲁棒性进行理论证明与实验验证. 由于 FLWB 方案具有传统水印的性质, 首先对 FLWB 方案的安全性进行理论证明. 有效性是为表明含水印模型的整体精度与无水印模型精度相近, 且含有所有用户的水印, 故与 FLWB 的功能保持性等价. FLWB 方案的鲁棒性通过常见的模型攻击测试进行验证.

3.1 安全性分析

假设模型 $\hat{M}$ 带有后门水印 $\mathcal{B}$, 则该模型的水印应无法被移除. 此外, 通过承诺方案的隐藏属性, 验证密钥不会向对手提供关于所用后门的任何有用信息, 而绑定属性确保任一非系统用户不能随意声称对该模型的所有权. 由于在构建方案时是从各用户的私有后门任务映射到水印算法, 故系统所需安全性将会提升, 意味着为达到破坏水印算法所需要的时间应比破坏后门时间更长. 在此前提下, 对第 2.2 节中所提水印性质进行安全性证明.

由于 FLWB 方案的输出是带有各用户水印的全局模型, 即最终各用户所获得的模型相同, 均为 $\hat{M}$, 故 FLWB 方案在正确性与功能保持性的证明与文献 [13] 相似, 均保证了在正常数据的分类任务上保持高精度, 且仅对后门数据上进行指定的误分类输出. 同时由于不可伪造性依赖于承诺方案的安全性, 且承诺方案在统计上的隐藏属性导致攻击者 $\mathcal{A}$ 无法通过承诺 c 来区分 VK 的不同, 所以 FLWB 方案的不可伪造性得以保证. 接下来对 FLWB 方案的不可移除性和非平凡所有权属性进行证明.

● 不可移除性. 若攻击者 $\mathcal{A}$ 无法在 PPT 时间 t 内消除模型 $\hat{M}$ 中的水印, 则表示 FLWB 算法具有不可移除性. 由于 FL 模型 M 的主要训练步骤 FedTrain 规定是在 PPT 时间 T 内完成的, 故在此假设若 $\mathcal{A}$ 无法在时间 t 内生成 $(1-\epsilon)$ 精度的不含水印的模型 Q, 则不可移除性成立. 其中 t 定义为比使用 FedTrain 所训练出相同精确的模型所需时间小得多的时间, 即 $t \ll T$. 为证明该性质的成立, 本文采用如下反证法.

假设存在敌手 $\mathcal{A}$ 能够破坏不可移除性, 则按照性质 3 不可移除性的定义, $\mathcal{A}$ 能够在给定输入 $\hat{M}, VK$ 时输出 $(1-\epsilon)$ 精度的模型 Q, 且对触发集 T 中元素至少有 $(1-\epsilon)$ 的概率被正确分类. $\mathcal{A}$ 可以通过 2 种方式进行破坏: (1) $\mathcal{A}$ 利用其他用户公开的验证密钥 VK(无法获得其他用户的标记密钥 MK), PPT 时间内指向性地删除 $\hat{M}$ 中对应的 MK; (2) 重新训练不含有水印的 $(1-\epsilon)$ 精度的模型. 由于 VK 的生成利用了承诺方案, 且满足统计意义上的隐藏属性, 故 $\mathcal{A}$ 无法通过 VK 区分不同的绑定信息, 故无法指向性的生成与 VK 所对应的 MK, 所以无法对水印进行指向性地删除操作. 对于 (2), 由于 $\mathcal{A}$ 仅拥有其私有数据 D_{Adv}, 而无法获取其他用户数据 $\bar{D}$, 故在无法通过 (1) 的方式进行攻击的情况下, $\mathcal{A}$ 只能采用与 $\hat{M}$ 相同的训练方式 FedTrain 进行训练, 以获得 $(1-\epsilon)$ 精度的模型 N, 满足 $\Pr\limits_{x\in D_{Adv}}[\text{InferPre}(x,N)=f(x)] \approx \Pr\limits_{x\in\bar{D}}[\text{InferPre}(x,M)=f(x)]$. 所以 (2) 的方式攻击成功需要时间 $t' \approx T \gg t$, 与假设相悖, 故 FLWB 方案的不可移除性得以保证.

● 非平凡所有权. 敌手 $\mathcal{A}$ 作为 FL 系统外部实体, 会声称对于模型 M 的所有权. 由于承诺方案的存在, $\mathcal{A}$ 无法在得到 M 后更改标记密钥 $\tilde{MK}$ 和验证密钥 $\tilde{VK}$, 所以必须事先在 $\tilde{T}$ 的 $(1-\epsilon)$ 的比例上输出指定的 $\tilde{T}_L$. FedKeyGen 算法在候选集中均匀随机挑选样本生成 T, 由于用户样本的私有存储性质, $\mathcal{A}$ 无法拥有与任一参与用户完全相同的训练数据集, 故所选触发集 $\tilde{T}$ 与 T 相交的概率可忽略不计 (详见第 1.3 节). 为简单处理, 假设 $\Pr\left[\tilde{T}\cap T=\varnothing\right]=1$,

则 $\tilde{T}$ 中元素可以在 $\bar{D}$ 的内部或外部. 设 $\tilde{T}$ 在 $\bar{D}$ 的内部元素个数为 $n_1=\left|\bar{D}\cap\tilde{T}\right|$, $\bar{D}$ 的外部元素个数为 $n_2=\left|\tilde{T}\right|-n_1$.

本文对敌手能力进行增强, 当 $\mathcal{A}$ 获取到模型 M 后, 其将 M 在 $x\in\bar{D}\cap\tilde{T}$ 上输出的错误标签并入 $\tilde{T}_L$. 根据联邦学习后门的定义 (详见第 1.3 节), M 需在 $\bar{D}\cap\tilde{T}$ 上以 $(1-\epsilon)n_1$ 的比例输出指定目标标签作为已被承诺方案承诺的标签, 然而 M 在 $\bar{D}$ 上表现为 $(1-\epsilon)$ 精度, 所以在 $\bar{D}\cap\tilde{T}$ 上仅存在 ϵn_1 的比例分类错误. 因为当 $\epsilon<0.5$ 时, $\epsilon n_1<(1-\epsilon)n_1$ 总是成立, 且 ϵ 的取值往往远小于 0.5, 故 $\tilde{T}$ 无法在 $\bar{D}$ 中取得.

当 $\tilde{T}$ 中元素可以在 $\bar{D}$ 的外部时, 即 $\tilde{T}\subset D\backslash\bar{D}$, 则根据第 1.2 节中对 FL 所做的假设: 如果输入是独立于 M 随机选择并且在 $\bar{D}$ 之外, 那么 M 将会在期望分类中出现 $(|L|-1)n_2/|L|$ 个元素出现分类错误. 可以得到当 $\epsilon<0.5$ 且 $L\geqslant 2$ 时, $\epsilon n_2<(|L|-1)n_2/|L|$.

故对于非平凡所有权属性有以下结论: 由于 $\epsilon n=\epsilon n_1+\epsilon n_2$, $\tilde{T}$ 的误差必须大于 ϵn 才能声称对于模型 M 的所有权, 与 $\tilde{T}$ 需保证 $(1-\epsilon)$ 精度 (后门水印的定义) 的条件相悖, 故 FLWB 方案的非平凡所有权得以保证.

综上所述, FLWB 方案满足第 2.2 节所定义的正确性、功能保持性、不可移除性、不可伪造性和非平凡所有权性. 因此, FLWB 方案被证明是安全的.

3.2 有效性与鲁棒性评估

3.2.1 实验设置

本节将通过实验验证 FLWB 方案的有效性及鲁棒性. 本文通过 TensorFlow 模拟 FL 系统, 系统环境如下: Windows 10、CPU 4 核、内存 8 GB、Python 3.6、TensorFlow 1.15.0. 实验使用的神经网络结构[22]见表 2.

本文所用数据集为 MNIST 和 Fashion-MNIST (https://github.com/zalandoresearch/fashion-mnist), 各数据集分别含有训练集 *TrS* 图像 50000 张, 测试 *TeS* 图像10000 张. 其余训练参数定义如表 3 所示.

表 2 神经网络结构图

网络层名称	参数设置
卷积层	5×5, 64
激活层	ReLU
卷积层	5×5, 64
激活层	ReLU
Dropout层	0.25
全连接层	128
激活层	ReLU
Dropout层	0.5
全连接层	10

表 3 FL 系统参数表

名称	参数设置
用户数量 N	10
每轮参与训练用户比例 C	1
优化算法	Adam
学习率 lr	0.001
本地训练批次大小B	50
本地迭代轮数E	5

由第 2.3 节分析可知, FL 水印可由后门任务得到, 故本文的标记密钥 MK 由训练任务中特定图像集合所构成的触发集来表示 (即在此标记密钥与触发集等价). 注意: 用户 i 从其本地训练集 TrS_i 中随机选择 mk_i, 所有 mk_i 在语义上构成 MK, 实际操作时并未集中收集.

在训练过程中, 本文采用从头训练的方式, 使用最小化交叉熵损失作为目标. 由于各用户在本地进行后门训练且标记密钥各不相同, 需保证每个本地后门模型的效果都能够映射到全局模型中, 同时因为每个用户的触发集样本数量远远小于主任务的样本数量, 故需强化触发集样本的训练过程. 本文通过将后门任务的训练过程与主任务训练过程解耦以增强触发集的效果, 各用户本地训练后门任务的过程详见算法 5.

3.2.2 有效性评估

FLWB 方案的有效性体现在不影响全局模型精度的前提下将用户水印进行嵌入, 与功能保持性保持等价. 所以各用户首先需要进行水印触发集的选择, 触发集的选择需要保持非平凡所有权属性中, 这要求 FL 系统的内部参与用户能够声称对于该模型的所有权, 而外部攻击者即使知道所用水印算法, 在不了解具体水印图片时也无法声称对该模型的所有权. 为充分满足该要求, 本文采用随机挑选的方式让各用户选择触发集, 并随机生成对应目标

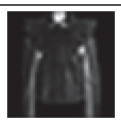
类, 之后各用户保证对该信息的私有化.

由于假设各用户所拥有的样本集是不相交的, 故上述采样方式保证了触发集的不相交. 所以即使部分用户泄露了其本地触发集的部分数据, 其他用户的触发集和该用户的其他触发集样本并未被泄露, 泄露的部分数据无法通过水印验证算法 Verify. 此外, 由于触发集样本和目标标签都是随机生成的, 使得基于反向传播反演样本的攻击方式变为极为困难. 表 4 和表 5 分别展示了 10 个用户在 MNIST 和 Fashion-MNIST 数据集上的触发集样本和目标标签示例.

表 4　各用户在 MNIST 数据集上的触发集样本及标签

用户ID	1	2	3	4	5
触发集样本					
源标签	1	0	9	5	4
目标标签	5	3	1	6	9
用户ID	6	7	8	9	10
触发集样本					
源标签	8	1	6	9	0
目标标签	0	7	8	4	2

表 5　各用户在 Fashion-MNIST 数据集上的触发集样本及标签

用户ID	1	2	3	4	5
触发集样本					
源标签	7	7	6	9	2
目标标签	4	2	1	5	6
用户ID	6	7	8	9	10
触发集样本					
源标签	6	3	0	5	1
目标标签	9	7	8	3	0

注: 0: T-shirt/top; 1: Trouser; 2: Pullover; 3: Dress; 4: Coat; 5: Sandal; 6: Shirt; 7: Sneaker; 8: Bag; 9: Ankle boot

FL-baseline-MNIST 和 FL-baseline-fMNIST 是基于表 3 参数在 MNIST 和 Fashion-MNIST 数据集上训练生成的联邦学习模型. FLWB 方案需要拥有功能保持性, 即要求含有水印的模型 FLWB 与不含水印的模型 FL-baseline 性能相似. 针对分类问题, 本文采用准确率作为评价标准. 由于各用户仅使用单样本作为标记密钥, 本文采用触发集成功个数作为评价指标, 当 Top-1 预测结果与目标标签相同时则记为该后门样本嵌入成功. 表 6 总结了含水印模型与不含水印模型在测试集和触发集上的准确率.

由表 6 可见, FLWB 方案准确率有轻微降低, 这是因为随机生成的目标标签对总体模型有一定的影响, 但总体

相差较小. 同时不含水印模型的触发集成功个数为 0, 而 FLWB 在规定的触发集中表现良好, 故 FLWB 方案的功能保持性得以保证.

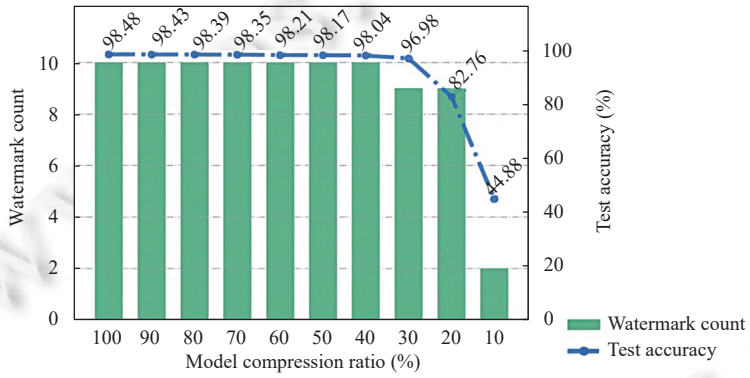
表 6 含水印模型与不含水印模型效果总结表

模型	测试集Top-1准确率 (%)	触发集成功个数
FL-baseline-MNIST	99.13	0
FLWB-MNIST	98.60	10
FL-baseline-fMNIST	90.88	0
FLWB-fMNIST	89.82	10

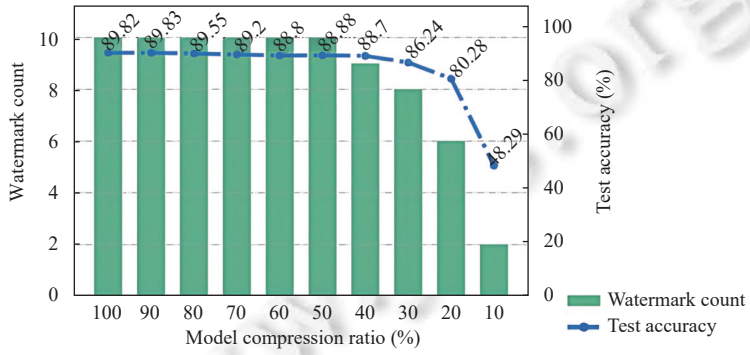
3.2.3 鲁棒性评估

为研究 FLWB 的鲁棒性, 需根据不可移除性的定义来研究已嵌入水印的模型在尝试去除水印时模型功能的变化, 否则直接将所有权重值置零能够完全去除水印但会严重损害模型性能. 本文主要考虑 2 种攻击方式下的模型效果: 基于 Top-K 的模型压缩攻击和模型微调 (fine-tuning) 攻击.

(1) 在联邦学习中, 基于 Top-K 的参数选择方法是常用的模型压缩方法[23], 能够在保证模型功能完好的情况下对模型参数进行压缩. 本文通过将神经网络分层, 仅保留每层中绝对值最大的前 K 个参数作为最终模型参数, 结果如图 3 所示, 其中横坐标为每层保存参数的百分比数.



(a) MNIST 数据集下的水印模型性能



(b) Fashion-MNIST 数据集下的水印模型性能

图 3 基于 Top-K 模型压缩攻击下水印模型性能图

由图 3 可知, 在 MNIST 的结果中, 当各层参数压缩到 40% 及以上时, 模型整体性能大致相同, 同时水印能够全部保留. 随着模型参数量的减少, 模型整体性能和水印都会产生一定影响. 当参数量压缩到原始参数量的 20% 时, 模型整体效果下降约 16%, 而全局水印能维持在 9 个. 在 Fashion-MNIST 的结果中, 各层压缩参数比在

50% 以上时, 模型整体性能下降不明显, 且水印能够全部保留. 当各层参数压缩到原来的 30% 时, 模型性能下降 3%, 水印能够保留 80%. 随着模型保留参数的减少, 当参数仅压缩到 20% 时, 水印能够保留超过半数, 但模型性能下降约 10% 超过了不可移除性定义的 ϵ; 当参数压缩到 10% 时模型性能损耗明显, 但依然能够保留 20% 的水印, 即 2 个参与用户的水印得以保留.

由上述 2 组实验可以得出结论, FLWB 方案对 Top-K 模型压缩攻击具有良好的鲁棒性. 考虑该种情况出现的原因, 可能是因为若想造成植入后门, 需要在不影响模型整体效果的情况下对该触发集样本起作用的部分参数进行适量增大, 而 Top-K 的模型压缩方式因为保留了绝对值比较大的参数, 也就保留了引起后门任务的参数.

(2) 微调方法是深度学习领域常用的提高模型精度的方法, 其能够针对特定数据集进行参数微调使得模型能更好地拟合该数据. 由于该方法仅需较少的计算资源和训练数据就能在目标数据集上取得更好的效果, 所以攻击者可以采取微调的方法来尝试对水印的消除. 本文采用 4 种不同的微调方法来验证 FLWB 方案的鲁棒性.

1) 微调最后一层参数 (fine-tune last layer, FTLL): 仅更新最后一层参数. 在设置中, 将其他层参数进行冻结, 而在微调时仅允许最后一层参数更新. 由于最后一层是输出层, 输出各样本对于不同类别的概率, 故可将此设置视为当喂入新输入特征而仅对输出微调的情况.

2) 微调所有层 (fine-tune all layers, FTAL): 更新所有层参数.

3) 重训练最后一层 (re-train last layer, RTLL): 将最后一层参数使用随机数初始化并重新训练, 其他层参数保持不变且冻结. 该方法主要为验证水印模型在存在噪音时的鲁棒性.

4) 重新训练所有层 (re-train all layers, RTAL): 使用随机数初始化最后一层参数, 其他层参数可以更新.

在使用上述 4 种微调方案时, 本文假设攻击者在 MNIST 和 Fashion-MNIST 的测试集中随机选择 3 000 个样本用以微调, 并且微调所用参数与训练过程参数相同 (除学习率 lr), 即训练轮数 $E=5$, 批大小 $B=50$, 学习算法为 Adam. 由于模型微调的学习率往往小于训练时的学习率大小, 故设置为 $lr=1\times10^{-4}$. 表 7 展示了含有水印模型在上述 4 种微调方法的攻击下的鲁棒性.

表 7　微调攻击下水印模型效果总结表

数据集	微调方法	测试集Top-1准确率 (%)	触发集成功个数 (个)
MNIST	FTLL	98.67	10
	FTAL	98.79	9
	RTLL	97.73	10
	RTAL	98.65	9
Fashion-MNIST	FTLL	90.04	10
	FTAL	90.22	9
	RTLL	87.80	9
	RTAL	90.04	9

结果表明, 所有微调方法下水印模型均保留有较高的测试集准确率, 并且后门水印保留情况良好. 综上所述, FLWB 方案具有良好的鲁棒性.

4 结　论

本文针对联邦学习中存在的模型产权保护问题, 利用传统后门任务中仅在少数样本误分类而保证模型主任务准确性的特征, 将每个用户的私有后门任务通过中心服务器聚合来映射到全局模型, 从而构建了一种有效的联邦学习水印方案. 本文对联邦学习水印进行密码学上的形式化描述, 并对联邦学习水印的性质进行定义, 同时为让全局模型尽可能包含所有参与用户的水印, 即降低各用户水印在全局模型中的冲突, 提出一种分步训练的方式, 让各用户在本地私有训练的过程中强化对触发集样本的训练, 从而将所有用户的水印映射到全局模型. 最后, 从理论和实验两方面证明了本文方案的安全性、有效性和鲁棒性. 由于更为复杂的网络层次对于触发集的选择和训练方式

有着更高的要求, 后续研究可以考虑通过对目标函数设置约束的方式对现有方案进行扩展, 以适应更复杂的深度学习网络结构. 另外, 提高联邦学习水印方案抵御重写攻击、歧义攻击等主动强攻击形式的能力仍是后续研究的重点和难点. 同时, 本文拟在未来工作中对水印选择、水印容量和模型结构的关系进行探讨, 进一步提高联邦学习模型的水印嵌入容量.

References:

[1] Yang Q, Liu Y, Chen TJ, Tong YX. Federated machine learning: Concept and applications. ACM Trans. on Intelligent Systems and Technology, 2019, 10(2): 12. [doi: 10.1145/3298981]

[2] Li T, Sahu AK, Talwalkar A, Smith V. Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 2020, 37(3): 50–60. [doi: 10.1109/MSP.2020.2975749]

[3] McMahan B, Moore E, Ramage D, Hampson S, Arcas BA. Communication-efficient learning of deep networks from decentralized data. In: Proc. of the 20th Int'l Conf. on Artificial Intelligence and Statistics. Fort Lauderdale: JMLR, 2017. 1273–1282.

[4] Li QB, Wen ZY, Wu ZM, Hu SX, Wang NB, Li Y, Liu X, He BS. A survey on federated learning systems: Vision, hype and reality for data privacy and protection. IEEE Trans. on Knowledge and Data Engineering, 2023, 35(4): 3347–3366. [doi: 10.1109/TKDE.2021.3124599]

[5] Deng YH, Lyu F, Ren J, Chen YC, Yang P, Zhou YZ, Zhang YX. FAIR: Quality-aware federated learning with precise user incentive and model aggregation. In: Proc. of the 2021 IEEE Conf. on Computer Communications. Vancouver: IEEE, 2021. 1–10. [doi: 10.1109/INFOCOM42981.2021.9488743]

[6] Tang M, Wong VWS. An incentive mechanism for cross-silo federated learning: A public goods perspective. In: Proc. of the 2021 IEEE Conf. on Computer Communications. 2021. Vancouver: IEEE. 2021. 1–10. [doi: 10.1109/INFOCOM42981.2021.9488705]

[7] Fredrikson M, Jha S, Ristenpart T. Model inversion attacks that exploit confidence information and basic countermeasures. In: Proc. of the 22nd ACM SIGSAC Conf. on Computer and Communications Security. Denver: ACM, 2015. 1322–1333. [doi: 10.1145/2810103.2813677]

[8] Hou RT, Xian HQ, Li J, Di GD. Graded reversible watermarking scheme for relational data. Ruan Jian Xue Bao/Journal of Software, 2020, 31(11): 3571–3587 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5812.htm [doi: 10.13328/j.cnki.jos.005812]

[9] Hu RW, Xiang SJ. Cover-lossless robust image watermarking against geometric deformations. IEEE Trans. on Image Processing, 2021, 30: 318–331. [doi: 10.1109/TIP.2020.3036727]

[10] Liang XY, Xiang SJ. Robust reversible audio watermarking based on high-order difference statistics. Signal Processing, 2020, 173: 107584. [doi: 10.1016/j.sigpro.2020.107584]

[11] Liu XY, Wang YF, Sun ZQ, Wang L, Zhao RC, Zhu YS, Zou BJ, Zhao YQ, Fang H. Robust and discriminative zero-watermark scheme based on invariant features and similarity-based retrieval to protect large-scale DIBR 3D videos. Information Sciences, 2021, 542: 263–285. [doi: 10.1016/j.ins.2020.06.066]

[12] Zhang YJ, Chen K, Zhou G, Lü PZ, Liu Y, Huang L. Research progress of neural networks watermarking technology. Journal of Computer Research and Development, 2021, 58(5): 964–976 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.2021.20200978]

[13] Adi Y, Baum C, Cissé M, Pinkas B, Keshet J. Turning your weakness into a strength: Watermarking deep neural networks by backdooring. arXiv:1802.04633, 2018.

[14] Wang TH, Kerschbaum F. RIGA: Covert and robust white-box watermarking of deep neural networks. In: Proc. of the 2021 Web Conf. Ljubljana: ACM, 2021. 993–1004. [doi: 10.1145/3442381.3450000]

[15] Szyller S, Atli BG, Marchal S, Asokan N. DAWN: Dynamic adversarial watermarking of neural networks. In: Proc. of the 29th ACM Int'l Conf. on Multimedia. ACM, 2021. 4417–4425. [doi: 10.1145/3474085.3475591]

[16] Fan LX, Ng KW, Chan CS. Rethinking deep neural network ownership verification: Embedding passports to defeat ambiguity attacks. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: NIPS, 2019. 4714–4723.

[17] Ong DS, Chan CS, Ng KW, Fan LX, Yang Q. Protecting intellectual property of generative adversarial networks from ambiguity attacks. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021. 3629–3638. [doi: 10.1109/CVPR46437.2021.00363]

[18] Li BW, Fan LX, Gu HL, Li J, Yang Q. FedIPR: Ownership verification for federated deep neural network models. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2023, 45(4): 4521–4536. [doi: 10.1109/TPAMI.2022.3195956]

[19] Gu TY, Liu K, Dolan-Gavitt B, Garg S. BadNets: Evaluating backdooring attacks on deep neural networks. IEEE Access, 2019, 7:

47230–47244. [doi: 10.1109/ACCESS.2019.2909068]

[20] Boneh D, Shaw J. Collusion-secure fingerprinting for digital data. IEEE Trans. on Information Theory, 1998, 44(5): 1897–1905. [doi: 10.1109/18.705568]

[21] Smart NP. Cryptography Made Simple. Switzerland: Springer, 2016. 430–450. [doi: 10.1007/978-3-319-21936-3]

[22] Bhagoji AN, Chakraborty S, Mittal P, Calo SB. Analyzing federated learning through an adversarial lens. In: Proc. of the 36th Int'l Conf. on Machine Learning. Long Beach: PMLR, 2019. 634–643.

[23] Dong Y, Hou W, Chen XJ, Zeng S. Efficient and secure federated learning based on secret sharing and gradients selection. Journal of Computer Research and Development, 2020, 57(10): 2241–2250 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.2020.20200463]

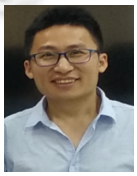
附中文参考文献:

[8] 侯瑞涛, 咸鹤群, 李京, 狄冠东. 分级可逆的关系数据水印方案. 软件学报, 2020, 31(11): 3571–3587. http://www.jos.org.cn/1000-9825/5812.htm [doi: 10.13328/j.cnki.jos.005812]

[12] 张颖君, 陈恺, 周赓, 吕培卓, 刘勇, 黄亮. 神经网络水印技术研究进展. 计算机研究与发展, 2021, 58(5): 964–976. [doi: 10.7544/issn1000-1239.2021.20200978]

[23] 董业, 侯炜, 陈小军, 曾帅. 基于秘密分享和梯度选择的高效安全联邦学习. 计算机研究与发展, 2020, 57(10): 2241–2250. [doi: 10.7544/issn1000-1239.2020.20200463]

李璇(1984－), 女, 博士, 副教授, 主要研究领域为云计算安全, 联邦学习, 隐私保护.



金彪(1985－), 男, 博士, 副教授, CCF 专业会员, 主要研究领域为隐私保护, 数据挖掘.

邓天鹏(1997－), 男, 博士生, 主要研究领域为深度学习安全, 联邦学习, 隐私保护.

林劼(1972－) , 男, 博士, 教授, 主要研究领域为机器学习, 生物信息学, 生物信息安全.

熊金波(1981－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为物联网安全与隐私保护, 人工智能安全, 网联自动驾驶安全.