计算机科学
COMPUTER SCIENCE

抗推理攻击的隐私增强联邦学习算法

赵宇豪, 陈思光, 苏健

引用本文

赵宇豪, 陈思光, 苏健. 抗推理攻击的隐私增强联邦学习算法[J]. 计算机科学, 2023, 50(9): 62-67.

ZHAO Yuhao, CHEN Siguang, SU Jian. Privacy-enhanced Federated Learning Algorithm Against Inference Attack [J]. Computer Science, 2023, 50(9): 62-67.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220700174

抗推理攻击的隐私增强联邦学习算法

赵宇豪[1]　陈思光[1]　苏　健[2]
1 南京邮电大学物联网学院　南京 210003
2 南京信息工程大学计算机学院　南京 210044
(zyh19981202@163.com)

摘　要　联邦学习在保证各分布式客户端训练数据不出本地的情况下,由中心服务器收集梯度协同训练全局网络模型,具有良好的性能与隐私保护优势。但研究表明,联邦学习存在梯度传递引起的数据隐私泄漏问题。针对现有安全联邦学习算法存在的模型学习效果差、计算开销大和防御攻击种类单一等问题,提出了一种抗推理攻击的隐私增强联邦学习算法。首先,构建了逆推得到的训练数据与训练数据距离最大化的优化问题,基于拟牛顿法求解该优化问题,获得具有抗推理攻击能力的新特征。其次,利用新特征生成梯度实现梯度重构,基于重构后的梯度更新网络模型参数,可提升网络模型的隐私保护能力。最后,仿真结果表明所提算法能够同时抵御两类推理攻击,并且相较于其他安全方案,所提算法在保护效果与收敛速度上更具优势。

关键词:联邦学习;推理攻击;隐私保护;梯度扰动
中图法分类号　TP393

Privacy-enhanced Federated Learning Algorithm Against Inference Attack

ZHAO Yuhao[1],CHEN Siguang[1] and SU Jian[2]
1 School of Internet of Things,Nanjing University of Posts and Telecommunications,Nanjing 210003,China
2 School of Computer Science,Nanjing University of Information Science and Technology,Nanjing 210044,China

Abstract　In federated learning,each distributed client does not need to transmit local training data,the central server jointly trains the global model by gradient collection,it has good performance and privacy protection advantages. However,it has been demonstrated that gradient transmission may lead to the privacy leakage problem in federated learning. Aiming at the existing problems of current secure federated learning algorithms,such as poor model learning effect,high computational cost,and single attack defense,this paper proposes a privacy-enhanced federated learning algorithm against inference attack. First,an optimization problem of maximizing the distance between the training data obtained by inversion and the training data is formulated. The optimization problem is solved based on the quasi-Newton method to obtain new features with anti-inference attack ability. Second,the gradient reconstruction is achieved by using new features to generate gradients. The model parameters are updated based on the reconstructed gradients,which can improve the privacy protection capability of the model. Finally,simulation results show that the proposed algorithm can resist two types of inference attacks simultaneously,and it has significant advantages in protection effect and convergence speed compared with other secure schemes.

Keywords　Federated learning,Inference attack,Privacy preservation,Gradient perturbation

1　引言

传统的机器学习以集中式学习为主,集中式学习需要终端将数据上传到服务器,服务器基于收集到的数据执行学习任务,然而收集数据需要消耗大量的通信资源且存在数据泄漏的风险。为解决上述问题,2017 年谷歌提出了联邦学习[1]的概念。联邦学习分成中心服务器和本地客户端两个部分,各分布式客户端无需上传本地数据,仅需利用本地数据训练本地网络模型,然后上传梯度或模型参数至服务器进行聚合,服务器将聚合后的全局网络模型发送给各分布式客户端训练,不断重复上述过程直至达到最大通信轮次[1-3]。文献[4]证明了在理想状况下,经联邦学习训练后的网络模型效果

到稿日期:2022-07-18　返修日期:2023-01-06
基金项目:国家自然科学基金(61971235);江苏省"333 高层次人才培养工程"资助;中国博士后科学基金(面上一等资助)(2018M630590);江苏省博士后科研资助计划(2021K501C);南京邮电大学"1311"人才计划
This work was supported by the National Natural Science Foundation of China(61971235),333 High-level Talents Training Project of Jiangsu Province,China Postdoctoral Science Foundation(2018M630590),Jiangsu Planned Projects for Postdoctoral Research Funds(2021K501C) and 1311 Talents Plan of NJUPT.
通信作者:陈思光(sgchen@njupt.edu.cn)

接近集中式学习甚至更优。

然而，联邦学习框架的分布式特性不足以全面保护客户端免受威胁。最近的研究表明，好奇的服务器或潜在的攻击者，根据本地网络模型训练过程中产生的梯度或客户端训练完成后的网络模型参数，能够推测出本地隐私数据，此类攻击模式被称为推理攻击。例如，Zhu 等[5]提出了基于二范式的深度泄露算法(Deep Leakage From Gradients，DLG)，首次证明了梯度信息能够泄漏隐私数据。类似地，Geiping 等[6]在 DLG 的基础上提出了基于余弦相似度的反转梯度攻击，进一步证明了梯度信息可导致隐私数据泄漏的严重性。此外，Wang 等[7]提出了一种新的推理攻击，该攻击将生成对抗网络与多任务鉴别器相结合，能够恢复出隐私数据并识别其来源。

近期关于联邦学习的研究表明[8]，目前主要通过 3 种技术达到隐私保护效果：差分隐私[9-11]、梯度压缩[5]和同态加密[12-13]。差分隐私通过噪声机制，在训练数据、梯度或网络模型参数上添加噪声，对隐私数据进行保护。例如，Wei 等[14]在网络模型训练过程中添加一定的噪声，以达到保护的效果，但是差分隐私存在添加噪声过多而严重影响网络模型准确率的问题[15]。文献[16]证明了梯度下降算法中大部分的梯度交换是多余的，因此可以使用梯度压缩技术对梯度信息进行裁剪压缩，即将绝对值过小的梯度裁剪为零，以降低隐私泄漏的风险。梯度压缩与差分隐私存在同样的问题，即平衡隐私保护和网络模型准确率的问题。同态加密提供了一种密码学的解决方案，在各分布式客户端上传网络模型参数前，预先对网络模型参数进行加密，按照加密对象的不同，同态加密分为部分同态加密[17]和全同态加密[18]。Zhang 等[19]通过权重参数的同态加密来保证数据的隐私性。但是作为一种密码学方法，同态加密需要消耗大量的计算资源，难以应用于实际的联邦学习场景中。

近期有极个别研究方案基于梯度导致数据泄漏的本质原因构建安全方案。例如，Sun 等[20]提出方案 Soteria，用于解决联邦学习隐私泄漏问题，Soteria 以最大程度模糊逆推得到的训练数据为目标将部分梯度置零。尽管 Soteria 对 DLG 等攻击具有较好的防御性，但是该方案无法有效抵御针对梯度置零的推理攻击。

针对上述一系列安全技术方案存在的问题，本文提出了一种抗推理攻击的隐私增强联邦学习算法。具体地，基于网络模型预测准确率与保护效果的综合性考虑，构建逆推得到的训练数据与训练数据距离最大化的优化问题。通过拟牛顿法求解该优化问题，获得具有抗推理攻击能力的新特征，实现特征重构。其次，构建梯度重构操作，即利用新特征生成的全连接层梯度，代替原全连接层梯度，获得具有抗推理攻击能力的新梯度，并使用新梯度更新网络模型参数，提升网络模型的隐私保护能力。最后，仿真结果表明，本文提出的抗推理攻击隐私增强联邦学习算法在提供强隐私保护的同时，相较于其他安全方案，模型准确率更高且收敛速度更快。

本文第 2 章介绍了系统模型；第 3 章介绍抗推理攻击隐私增强联邦学习算法的相关定义及求解过程；第 4 章对算法性能与安全性进行分析；第 5 章为仿真结果及具体分析；最后总结全文并展望未来。

2 系统模型

联邦学习保证各分布式客户端训练数据不出本地，仅在中间阶段交换训练参数的情况下协同训练全局网络模型，避免了因训练数据收集产生的通信开销与安全隐患。虽然联邦学习不直接交换训练数据，相比传统的机器学习有更高的安全保障，但是该学习模式需要交换大量参数用于协同训练。经研究证实，好奇的服务器或潜在的攻击者可以通过模型参数或梯度推测出客户端的训练数据，带来隐私数据泄漏的风险。为解决现有联邦学习存在的问题，本文构建了抗推理攻击的隐私增强联邦学习模型，如图 1 所示。

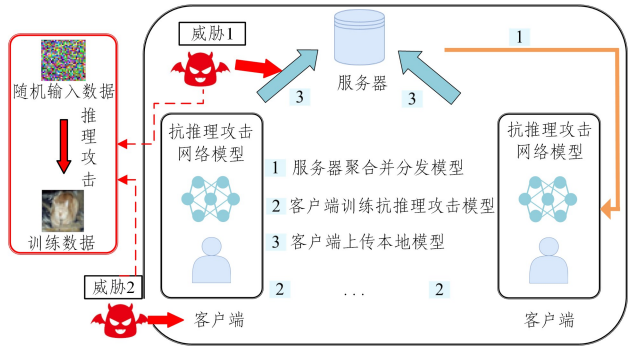


图 1 抗推理攻击的隐私增强联邦学习模型

Fig. 1 Privacy-enhanced federated learning model against inference attacks

本文所构建的联邦学习场景中假设主要包含两种推理攻击威胁，威胁 1 是能够窃听客户端与服务器之间通信链路的窃听者或诚实且好奇的服务器，根据窃听或直接获取到的网络模型参数进行推理攻击，恢复客户端的训练数据；威胁 2 发生于客户端训练网络模型的过程中，潜在的攻击者多为系统漏洞或携带病毒的程序，其根据捕获或控制得到的梯度进行推理攻击，恢复客户端的训练数据。上述两种威胁都需要通过有限的迭代来优化随机输入数据，使得输入数据生成的梯度近似于泄漏的梯度，优化后的输入数据即为生成的近似训练数据。

本文构建的抗推理攻击的隐私增强联邦学习模型须能够有效抵御上述两类威胁。该模型的设计主要包含两部分内容：

1)若是首轮联邦学习，则服务器将初始化网络模型发送至客户端，否则通过联邦平均算法聚合客户端上传的网络模型参数，并将聚合后的网络模型参数发送至各客户端，由服务器协调完成多轮联邦学习，直至获得最终的全局最优模型。

2)从上述的描述中可以看出联邦学习极易受到安全威胁，因此，要构建一个具备较强隐私保护能力的联邦学习模式，需要在保证网络模型准确率的前提下，使得攻击逆推得到的输入数据与训练数据的差距足够大。因为神经网络全连接层的梯度与参数是决定逆推得到的输入数据与训练数据相似性的关键，所以本文设计了一种全连接层梯度扰动方法以降低被攻击的风险，即通过有限的迭代来优化随机新特征，使得新特征近似于经卷积层提取后的特征，且新特征逆映射到网络模型的输入数据与训练数据差距足够大。优化后的新特征

会生成扰动梯度，该梯度具备抗推理攻击能力，并可代替原全连接层的梯度。使用该梯度更新网络模型参数，随着本地训练迭代次数的增加，由网络模型参数逆推得到的输入数据与训练数据的差距将不断增大，因此联邦学习的隐私保护能力也将不断提升。

3 抗推理攻击的隐私增强联邦学习算法

在传统的推理攻击场景中，攻击者基于所窃听或感知到的模型参数或梯度，能够逆推得到近似的训练数据，记为逆推输入数据 X'。目标函数 D 定义为逆推输入数据 X' 生成的梯度与泄漏梯度之间的二范式距离，推理攻击通过优化器优化目标函数 D，从而更新逆推输入数据 X'。更新的方式为：

$$\begin{aligned}(X',Y')&=\arg\min_{(X',Y')} D\\&=\arg\min_{(X',Y')} \| \nabla W'-\nabla W \|_2\\&=\arg\min_{(X',Y')} \| \frac{\partial \| f(X')-Y' \|_2}{\partial W}-\nabla W \|_2\end{aligned} \tag{1}$$

其中，X' 与 Y' 分别表示逆推输入数据与标签；W 表示网络模型参数；$\nabla W'$ 表示逆推输入数据 X' 生成的梯度；∇W 表示泄漏的梯度；$f(\cdot)$ 表示模型输入映射到目标结果的函数。在优化器的作用下，梯度 $\nabla W'$ 的大小将不断逼近梯度 ∇W 的大小，使得逆推输入数据 X' 不断逼近原始训练数据。优化后的逆推输入数据 X' 即为生成的近似训练数据，与训练数据相似性越高，推理攻击的效果就越好。

基于联邦学习存在的推理威胁，目前有一系列安全技术被提出来解决此类问题。目前比较流行的联邦学习安全技术包括差分隐私、同态加密和梯度压缩等，但是这些方法存在网络模型准确率较低或计算开销大的问题。例如，差分隐私的原理是在训练数据、梯度或网络模型参数上添加噪声，从而在一定程度上扰动数据，增加了推理攻击的难度。然而，过多的噪声会影响网络模型预测的准确率，这是差分隐私不可避免的困境。梯度压缩的原理是将绝对值较小的梯度置零，该方法的一个主要参数为 $mask$，$mask$ 决定梯度是否被置零，可以表示为：

$$mask=\begin{cases}0, & |\nabla W|<threshold\\ \nabla W, & |\nabla W|\geqslant threshold\end{cases} \tag{2}$$

其中，$|\nabla W|$ 表示梯度的绝对值，$threshold$ 表示裁剪率，低于该值的梯度将被置零，高于该值的梯度则不变。

基于式(2)，梯度压缩将部分梯度置零，使得在执行推理攻击的过程中，目标函数 D 被错误计算，从而间接导致在后续攻击的每一轮迭代过程中，逆推输入数据 X' 内的所有数据参数都被错误优化，最终无法准确地恢复训练数据。然而，梯度压缩方法无法有效抵御部分针对梯度置零的推理攻击，在针对梯度置零发起的推理攻击中，目标函数 D 将遍历值为零的梯度 ∇W，且不计算这些梯度 ∇W 与梯度 $\nabla W'$ 之间的二范式距离，即不优化被置零梯度对应的数据参数。此场景下的目标函数 D 定义为：

$$D=\begin{cases}\| \nabla W'-\nabla W \|_2, & \nabla W\neq 0\\ 0, & \nabla W=0\end{cases} \tag{3}$$

因此在后续攻击的每一轮迭代过程中，除被置零梯度对应的数据参数无法优化外，其余数据参数都能够被优化，从而极大地削减了梯度压缩的保护效果，最终恢复出近似的训练数据。

针对上述一系列安全技术存在的问题，本文设计了抗推理攻击的隐私增强联邦学习算法。该算法包含两个核心步骤：特征重构与梯度重构。

特征重构的对象是某个训练数据经卷积层提取后的特征，重构的需求被规划成两个优化问题：

1)新特征 r' 与原特征 r 的距离最小化问题。需要尽可能减小新特征 r' 与原特征 r 的二范式距离，保证网络模型的预测准确率接近无保护状态下的准确率。具体优化问题构建如下：

$$\min_{x'} \ \| r-r' \|_2 \tag{4}$$

2)逆推输入数据 X' 与训练数据 X 的距离最大化问题。需要尽可能增大逆推输入数据 X' 与训练数据 X 的二范式距离，保证推理攻击恢复不出训练数据。具体优化问题构建如下：

$$\max_{x'} \ \| X-X' \|_2 \tag{5}$$

根据拉格朗日定理可知，逆推输入数据 X' 与训练数据 X 存在如下近似相等关系：

$$\| X-X' \|_2 \approx \| (r-r')g_X'(r) \|_2 \tag{6}$$

其中，$g_X'(\cdot)$ 表示特征逆映射到模型输入的导函数。

结合式(4)－式(6)，可以得到最终的优化问题：

$$\min_{r'} \| r-r' \|_2+\alpha \left\| \frac{1}{(r-r')g_X'(r)} \right\|_2 \tag{7}$$

$$\text{s.t. } \varepsilon_1\leqslant \| r-r' \|_2\leqslant\varepsilon_2 \tag{7a}$$

$$\alpha>0 \tag{7b}$$

其中，ε_1 和 ε_2 表示新特征 r' 与原特征 r 之间二范式距离的取值范围；α 表示保护等级，其值越大，隐私保护效果越好。

基于拟牛顿法可求解上述优化问题。首先，设置合适的保护等级 α；其次，基于式(7)、式(7a)和式(7b)，在优化器的作用下，不断更新参数 r'，得到新特征 r' 的最优解，特征重构完成。

梯度重构的对象是使用某个训练数据训练网络模型时生成的全连接层梯度。传统的网络模型梯度由两部分构成，分别是卷积层的梯度 ∇W_{conv} 与全连接层的梯度 ∇W_{full}。卷积层的梯度 ∇W_{conv} 可以表示为：

$$\nabla W_{\text{conv}}=\frac{\partial \| g(X)-r \|_2}{\partial W_{\text{conv}}} \tag{8}$$

其中，$g(\cdot)$ 表示模型输入映射到特征的函数。全连接层的梯度 ∇W_{full} 可以表示为：

$$\nabla W_{\text{full}}=\frac{\partial \| F(r)-Y \|_2}{\partial W_{\text{full}}} \tag{9}$$

其中，$F(\cdot)$ 表示特征映射到目标结果的函数。

与基于式(9)的梯度生成方式不同，本文通过新特征 r'，生成具有抗推理攻击能力的新全连接层梯度 $\nabla W'_{\text{full}}$，具体可以表示为：

$$\nabla W'_{\text{full}}=\frac{\partial \| F(r')-Y \|_2}{\partial W'_{\text{full}}} \tag{10}$$

结合式(8)与式(10)，以新全连接层梯度 $\nabla W'_{\text{full}}$ 代替原全连接层梯度 ∇W_{full}，可以得到新网络模型梯度 $\nabla W'$，即梯度重构完成。新网络模型的梯度 $\nabla W'$ 可以表示为：

$$\nabla W'=\{\nabla W_{\text{conv}},\nabla W'_{\text{full}}\} \tag{11}$$

为了能更详细地了解本文所提的抗推理攻击隐私增强

联邦学习算法，以服务器第 e 轮通信轮次为例，详细介绍抗推理攻击的隐私增强联邦学习新模式。

首先，服务器发送全局模型参数 $W(e)$ 至各分布式客户端；其次，客户端在每一轮本地迭代的训练过程中，基于式(7)、式(7a)和式(7b)，执行特征重构操作，获得新特征 r'；然后，基于式(8)、式(10)和式(11)，执行梯度重构操作，获得新网络模型的梯度 $\nabla W'$，其包含原卷积层梯度 ∇W_{conv} 与新全连接层梯度 $\nabla W'_{full}$。进而使用梯度下降算法，更新卷积层的参数，具体可以表示为：

$$W_{conv}(e):=W_{conv}(e)-\eta\nabla W_{conv} \tag{12}$$

其中，$W_{conv}(e)$ 表示网络模型卷积层的参数；η 表示梯度下降的学习率。同时更新全连接层的参数，使其具备抗推理攻击能力，具体可以表示为：

$$W_{full}(e):=W_{full}(e)-\eta\nabla W'_{full} \tag{13}$$

其中，$W_{full}(e)$ 表示网络模型全连接层的参数。

当客户端达到最大迭代次数时，会上传训练后的网络模型参数 $W(e)_{over}$ 至服务器。当服务器收到所有客户端上传的网络模型参数 $W(e)_{over}$ 后，会计算各客户端网络模型的增量 $\Delta W(e)$，具体可以表示为：

$$\Delta W(e)=W(e)_{over}-W(e) \tag{14}$$

基于式(14)，服务器计算所有客户端网络模型的增量，进而更新全局网络模型参数 $W(e)$，具体可以表示为：

$$W(e+1)=W(e)+\frac{1}{N}\sum_{i=1}^{N}\Delta W^i(e) \tag{15}$$

其中，N 表示分布式客户端的数目，$\Delta W^i(e)$ 表示第 i 个客户端网络模型的增量。

基于式(14)和式(15)，服务器更新得到具有抗推理攻击能力的网络模型参数 $W(e+1)$，从而完成第 e 轮通信轮次的更新。若未达到最大通信轮次，服务器继续将网络模型参数 $W(e+1)$ 传送至各客户端进行训练，直至达到最大通信轮次。为便于理解上述执行流程，将此求解过程描述为算法 1 的形式。

算法 1 抗推理攻击的隐私增强联邦学习算法

输入：保护等级 α，随机新特征 r′，特征重构的优化迭代次数 K

输出：全局最优模型参数 W(E)

BEGIN

服务器：

1. 初始化网络模型参数 W(e)；
2. FOR e IN 通信轮次 E DO
3. 发送网络模型参数 W(e)至各分布式客户端；
4. 接收所有客户端上传的网络模型参数 W(e)$_{over}$；
5. 基于式(14)计算各网络模型的增量 ΔW(e)；
6. 基于式(15)更新全局网络模型参数 W(e)；
7. END FOR
8. 获得具有抗推理攻击能力的全局最优模型参数 W(E).

客户端：

9. FOR t IN 本地迭代次数 T DO
10. 初始化 α 和新特征 r′；
11. FOR k in 1,…,K DO
12. 基于式(7)、式(7a)和式(7b)优化新特征 r′；
13. END FOR
14. 基于式(8)、式(10)和式(11)计算新网络模型的梯度 ∇W′；
15. 基于式(12)和式(13)更新网络模型参数 W(e)；
16. END FOR
17. 获得更新后的网络模型参数 W(e)$_{over}$并上传至服务器。
18. END

4 算法性能与安全性分析

4.1 复杂度分析

相较于传统的联邦学习，本文算法复杂度的增加主要来源于特征重构优化问题的求解，即式(7)优化问题的求解。该优化问题的求解复杂度与训练数据批量大小有关，而真实场景下训练数据批量大小通常较小，即该优化问题的求解复杂度远低于联邦学习训练的复杂度，因此本文算法在获得较好隐私保护效果的同时，并没有显著增加联邦学习训练的复杂度。

4.2 安全性分析

从数据层面看，各分布式客户端训练数据不出本地，仅在中间阶段交换训练参数的情况下协同训练全局网络模型，使训练数据具有一定的安全性。从模型层面看，推理攻击通过泄漏的模型梯度来复原训练数据，尤其是全连接层的梯度，因此本文算法对全连接层梯度加以保护。考虑到全连接层的梯度由特征生成，且特征与训练数据之间存在映射关系，因此本文算法以最大化训练数据与复原后训练数据之间的二范式距离为目标，生成新的特征，由该特征得到的全连接层梯度，能够提升网络模型的隐私性。具体保护效果见第 5 章仿真结果。

5 仿真结果

本章通过仿真实验来评估所提抗推理攻击隐私增强联邦学习算法的有效性，并与一些经典的安全方案进行对比。本仿真样本集为 70 000 张 MNIST 和 60 000 张 CIFAR10 数据集。为了表明本文算法的稳定性，考虑了本文算法在 MNIST 数据集上进行独立同分布(Independent Identically Distributed，IID)与非独立同分布 Non-IID 的实验。

设定参与训练的联邦学习服务器个数为 1，通信轮次(Communication rounds)上限为 500，聚合算法为联邦平均算法，其学习率为 0.05；参与训练的分布式客户端个数为 10，本地训练迭代次数为 10，本地训练使用随机梯度下降算法，其学习率为 0.01，参数 *batch* 为 32。

在本仿真实验中，推理攻击包含两类：第一类为文献[5]所提出的推理攻击；第二类是针对梯度置零的推理攻击。若无特殊说明，则默认使用第一类推理攻击。均方差(Mean Squared Error，MSE)定义为逆推得到的近似图像与原始图像之间的均方差，Min-MSE 定义为攻击迭代过程中最小的 MSE 值，以 Min-MSE 作为评估保护效果的参数。

图 2 给出了在攻击迭代次数 $Iters=300$ 的设置下，抗推理攻击的隐私增强联邦学习算法在不同保护等级 α 下的保护效果随着 *Iters* 的变化情况。从该图可以看出，在 $\alpha=0.01$，$\alpha=0.05$ 的设置下，逆推得到的图像并没有随 *Iters* 的增加而清晰，表明了本文算法的有效性。同时可以发现，在 $\alpha=0.001$，$\alpha=0.005$ 的设置下，随着 *Iters* 的增加，逆推得到的图像渐渐近似于原始图像。这是因为 α 的大小会影响保护效果的强弱。当 $\alpha=0.001$，$\alpha=0.005$ 时，α 的值过小，不能产生具有充分抗推理攻击能力的新特征，直接影响了保护效果。因此，在后续实验中，将本文算法的保护等级 α 设置为 0.05。

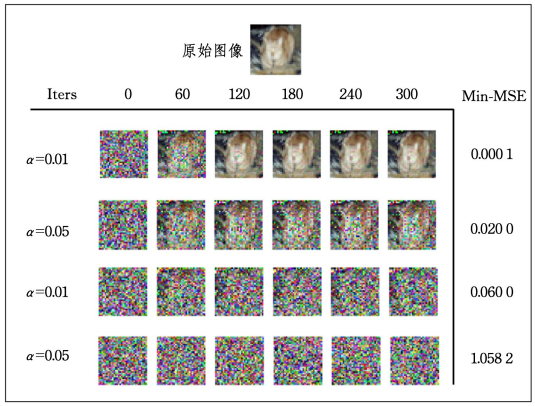


图 2 不同保护等级下保护效果随迭代次数的变化

Fig. 2 Variation of protection effects with iteration times under different protection levels

在后续仿真实验中，本文算法与现有的 3 种安全联邦学习模型 Soteria against model inversion attack in FL (Soteria)[20]，Federated client differential privacy (Fed-CDP)[14]，Federated learning based on adaptive compression (FL-AC)[21] 进行对比，对比的联邦学习模型中训练数据的分布为 Non-IID。将 Soteria 的裁剪率设置为 60%，将 FL-AC 的梯度压缩裁剪率设置为 25%，将 Fed-CDP 的噪声大小设置为 1×10^{-3}。

图 3 给出了在两类推理攻击 *Iters*=300 的设置下，保护效果随着 *Iters* 的变化情况。

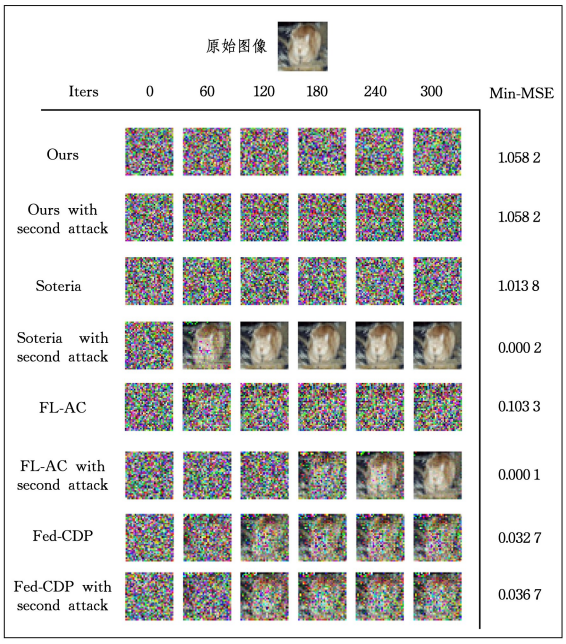


图 3 两类攻击下不同方案保护效果随迭代次数的变化

Fig. 3 Variation of protection effects of different schemes under two types of attacks

该图 3 可以看出，两类推理攻击在面对本文算法时，逆推得到的图像并不随 *Iters* 的增加而清晰，表明了本文算法的有效性。同时可以发现，第二类推理攻击在面对 Soteria 和 FL-AC 时，逆推得到的图像渐渐近似于原始图像，且 Fed-CDP 在两种推理攻击下保护效果较差。表明 Soteria 和 FL-AC 无法有效抵御第二类推理攻击，Fed-CDP 无法有效抵御两类推理攻击。本文算法能够同时抵御这两类推理攻击，是因为本文算法执行梯度重构操作，生成了能够充分抗推理攻击的新梯度，而不是仅仅将部分梯度置零。

图 4 给出了在各批次随机选取 10 个 MNIST 数据样本的场景下，各批次保护效果的平均 Min-MSE。本文算法包含 IID 与 Non-IID 两种情况，其中无标记的紫色粗直线表示恢复的图像是否能被人眼识别的边界。从该图可以看出，Fed-CDP 下的 Min-MSE 小于 0.2，表明该方案无法有效抵御推理攻击，这是因为添加的噪声值过小。同时可以发现，无论是在 IID 还是 Non-IID 数据分布情况下，整体上本文算法的 Min-MSE 比其他安全方案更高，表明在本文算法的保护下，逆推得到的近似图像与原始图像之间的均方差更大，图像保护的效果越好。这是因为本文算法执行特征重构操作，其优化目标之一为最大化逆推得到的图像与原始图像之间的二范式距离，保证了逆推得到的近似图像无法被准确识别。

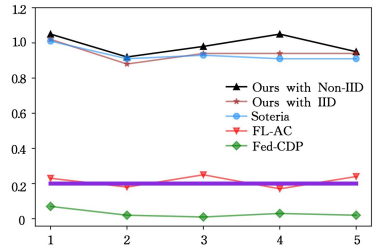


图 4 不同批次图像下不同方案保护效果的对比(电子版为彩图)

Fig. 4 Comparison of protection effect of different schemes under different batch images

图 5 给出了不同安全方案平均损失(Loss)随 Communication rounds 的变化曲线以及本文算法在 Non-IID 下的网络模型准确率(Accuracy)随 Communication rounds 的变化曲线(本文算法的 Loss 变化曲线包含 IID 与 Non-IID 两种情况)。从图中可以看出，本文算法的 Accuracy 高于 90%，表明其具有较高的准确率。同时可以看出，无论是在 IID 还是 Non-IID 数据分布情况下，本文算法的收敛速度都仅次于无保护状态下的网络模型，且在 *Communication rounds*<200 内 Loss 接近于 0，表明本文算法能以较快的速度收敛于更低的 Loss 值。这是因为本文算法以最大化逆推得到的图像与原始图像之间的二范式距离、最小化新特征与原特征的二范式距离为目标，执行特征重构操作，生成能够充分抗推理攻击的新特征，同时保持高准确率，接近无保护状态下的收敛效果。

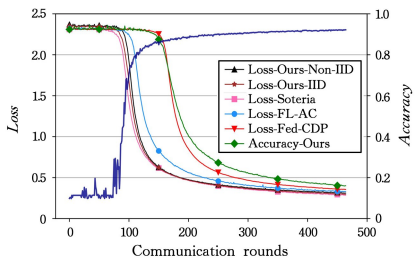


图 5 平均损失变化曲线与准确率变化曲线

Fig. 5 Average loss changing and accuracy changing curves

结束语 为了在提供强隐私保护的同时保证模型性能，本文提出了一种抗推理攻击的隐私增强联邦学习算法。构建了特征重构操作，以获得具有抗推理攻击能力的新特征。

新特征生成的梯度同样具有抗推理攻击能力，利用此梯度更新网络模型参数，最终提升网络模型的隐私保护能力。最后，仿真结果表明，本文提出的算法能够在保证网络模型高准确率与快速收敛的同时，有效抵御两类推理攻击。本文方法虽然提升了联邦学习的隐私保护能力，但并不能抵御其他类型的联邦学习攻击。在未来的工作中，将致力于设计更加安全的模型，使联邦学习能够抵御更多类型的攻击。

参 考 文 献

[1] MCMAHAN B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data [C]// Proceedings of the 20th International Conference on Artificial Intelligence and Statistics(AISTATS). 2016: 1273-1282.

[2] YANG Q, LIU Y, CHEN T, et al. Federated machine learning: Conceptand applications [J]. ACM Transactions on Intelligent Systems and Technology, 2019, 10(2): 1-19.

[3] BONAWITZ K, EICHNER H, GRIESKAMP W, et al. Towards federated learning at scale: System design [C]// Proceedings of Machine Learning and Systems(MLSys). 2019: 374-388.

[4] LI T, SAHU A K, TALWALKAR A, et al. Federated learning: Challenges, methods, and future directions [J]. IEEE Signal Processing Magazine, 2020, 37(3): 50-60.

[5] ZHU L, LIU Z, HAN S. Deep leakage from gradients [C]// Proceedings of Advances in Neural Information Processing Systems (NIPS). 2019: 17-31.

[6] GEIPING J, BAUERMEISTER H, DRÖGE H, et al. Inverting gradients-how easy is it to break privacy in federated learning? [C]// Proceedings of Advances in Neural Information Processing Systems(NIPS). 2020: 16937-16947.

[7] WANG Z, SONG M, ZHANG Z, et al. Beyond inferring class representatives: User-level privacy leakage from federated learning[C]// Proceedings of IEEE International Conference on Computer Communications(INFOCOM). 2019: 2512-2520.

[8] LIU J, MENG X. Survey on Privacy-Preserving Machine Learning[J]. Journal of Computer Research and Development, 2020, 57(2): 346-362.

[9] WEI K, LI J, DING M, et al. Federated learning with differential privacy: Algorithms and performance analysis [J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 3454-3469.

[10] MCMAHAN H B, RAMAGE D, TALWAR K, et al. Learning differentially private recurrent language models [C]// Proceedings of International Conference on Learning Representations (ICLR). 2018: 171-182.

[11] TRUEX S, LIU L, CHOW K H, et al. LDP-Fed: Federated learning with local differential privacy [C]// Proceedings of the Third ACM International Workshop on Edge Systems (EdgeSys). 2020: 61-66.

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

[13] LIU Y, KANG Y, XING C, et al. A secure federated transfer learning framework[J]. IEEE Intelligent Systems, 2020, 35(4): 70-82.

[14] WEI W, LIU L, WUT Y, et al. Gradient-leakage resilient federated learning [C]// Proceedings of the 41st IEEE International Conference on Distributed Computing Systems(ICDCS). 2021: 797-807.

[15] WU N, FAROKHI F, SMITH D, et al. The value of collaboration in convex machine learning with differential privacy [C]// Proceedings of IEEE Symposium on Security and Privacy(SP). 2020: 304-317.

[16] LIN Y, HAN S, MAO H, et al. Deep gradient compression: Reducing the communication bandwidth for distributed training [C]// Proceedings of International Conference on Learning Representations(ICLR). 2017: 1-12.

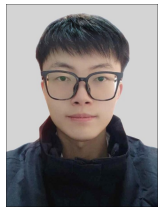
[17] MARTINS P, SOUSA L, MARIANO A. A survey on fully homomorphic encryption: An engineering perspective [J]. ACM Computing Surveys, 2017, 50(6): 1-33.

[18] ACAR A, AKSU H, ULUAGAC A S, et al. A survey on homomorphic encryption schemes: Theory and implementation [J]. ACM Computing Surveys, 2018, 51(4): 1-35.

[19] ZHANG Z, FU Y, HE N, GAO T. Research on Federated Deep Neural Network Model for Data Privacy Preserving[J]. Acta Automatica Sinica, 2022, 48(5): 1273-1284.

[20] SUN J, LI A, WANG B, et al. Soteria: Provable defense against privacy leakage in federated learning from representation perspective [C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). 2021: 9311-9319.

[21] JIANG B, LI J, WANG H, et al. Privacy-Preserving federated learning for industrial edge computing via hybrid differential privacy and adaptive compression [J]. IEEE Transactions on Industrial Informatics, 2023, 19(2): 1136-1144.



ZHAO Yuhao, born in 1998, postgraduate. His main research interest is federated learning.

CHEN Siguang, born in 1984, Ph.D, professor. His main research interests include edge intelligence and AIoT.

（责任编辑：何杨）