基于对比训练的联邦学习后门防御方法

张佳乐[1,2]，朱诚诚[1]，成翔[1,2]，孙小兵[1]，陈兵[3]
（1. 扬州大学信息工程学院，江苏 扬州 225127；2. 中国民航大学民航飞联网重点实验室，天津 300300；
3. 南京航空航天大学计算机科学与技术学院，江苏 南京 211106）

摘 要：针对现有联邦学习后门防御方法不能实现对模型已嵌入后门特征的有效清除同时会降低主任务准确率的问题，提出了一种基于对比训练的联邦学习后门防御方法 ContraFL。利用对比训练来破坏后门样本在特征空间中的聚类过程，使联邦学习全局模型分类结果与后门触发器特征无关。具体而言，服务器通过执行触发器生成算法构造生成器池，以还原全局模型训练样本中可能存在的后门触发器；进而，服务器将触发器生成器池下发给各参与方，各参与方将生成的后门触发器添加至本地样本，以实现后门数据增强，最终通过对比训练有效消除后门攻击的负面影响。实验结果表明，ContraFL 能够有效防御联邦学习中的多种后门攻击，且效果优于现有防御方法。

关键词：联邦学习；后门攻击；对比训练；触发器；后门防御
中图分类号：TP393
文献标志码：A
DOI: 10.11959/j.issn.1000−436x.2024063

Backdoor defense method in federated learning based on contrastive training

ZHANG Jiale[1,2], ZHU Chengcheng[1], CHENG Xiang[1,2], SUN Xiaobing[1], CHEN Bing[3]
1. School of Information Engineering, Yangzhou University, Yangzhou 225127, China
2. Key Laboratory of Flying Internet, Civil Aviation University of China, Tianjin 300300, China
3. College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China

Abstract: In response to the inadequacy of existing defense methods for backdoor attacks in federated learning to effectively remove embedded backdoor features from models, while simultaneously reducing the accuracy of the primary task, a federated learning backdoor defense method called ContraFL was proposed, which utilized contrastive training to disrupt the clustering process of backdoor samples in the feature space, thereby rendering the global model classifications in federated learning independent of the backdoor trigger features. Specifically, on the server side, a trigger generation algorithm was developed to construct a generator pool to restore potential backdoor triggers in the training samples of the global model. Consequently, the trigger generator pool was distributed to the participants by the server, where each participant added the generated backdoor triggers to their local samples to achieve backdoor data augmentation. Experimental results demonstrate that ContraFL effectively defends against various backdoor attacks in federated learning, outperforming existing defense methods.

Keywords: federated learning, backdoor attack, contrastive training, trigger, backdoor defense

收稿日期：2023−11−13；**修回日期：**2024−02−19
通信作者：朱诚诚，MX120220554@stu.yzu.edu.cn
基金项目：国家自然科学基金资助项目（No.62206238）；江苏省基础研究计划自然科学基金资助项目（No.BK20220562）；江苏省高等学校基础科学（自然科学）研究基金资助项目（No.22KJB520010）；中国博士后科学基金资助项目（No.2023M732985）；中国民航大学民航飞联网重点实验室开放基金资助项目（No.MHFLW202304）；江苏省研究生科研创新计划基金资助项目（No.KYCX23_3534）
Foundation Items: The National Natural Science Foundation of China (No.62206238), The Basic Research Program Natural Science Foundation of Jiangsu Province (No.BK20220562), The Natural Science Foundation of Jiangsu Higher Education Institutions (No.22KJB520010), The China Postdoctoral Science Foundation (No.2023M732985), The Open Fund for the Key Laboratory of Flying Internet at Civil Aviation University of China (No.MHFLW202304), The Postgraduate Research and Practice Innovation Program of Jiangsu Province (No.KYCX23_3534)

0　引言

近年来，随着物联网、边缘计算、5G 等技术的不断发展及用户终端数量的爆炸式增长，传统云计算架构下的集中式机器学习模型由于具备高时延、高并发、弱隐私保护等缺陷，已经逐渐演化成为能够支撑边缘智能化应用的分布式联邦学习架构[1-3]。联邦学习技术能够使机器学习模型以分布式、本地化的训练方式得到不断改善和优化，是近年来解决机器学习中数据孤岛问题和实现隐私保护的重要手段。因此，将新型联邦学习框架应用于智能数据分析已经成为一个重要的研究方向，国内外各大企业也相应推出多种联邦学习应用框架，如 Google 和 TensorFlow 官方研发的 TFF[4]、微众银行的 FATE[5]等。可以说，联邦学习已呈现出极具实用性的发展潜力，有关联邦学习的研究方向也被国内外学者广泛关注[6-9]。

然而，标准联邦学习算法极易遭受不可信参与方发起的对抗性攻击[10-11]，攻击方式通常是以非正常的本地更新（恶意梯度）来影响全局模型的准确性，如后门攻击[12]、投毒攻击[13-14]和对抗样本攻击[15]等。其中，后门攻击是一种面向深度神经网络（DNN，deep neural network）的典型对抗性攻击，Gu 等[16]首次探索了对 DNN 的后门攻击方法 BadNets，其中，攻击方试图在 DNN 中嵌入隐藏的后门模式，使被攻击的 DNN 在良性样本上正常分类，但当输入样本激活了隐藏的后门模式时，模型将输出攻击方指定标签。目前，常用的后门触发器类型包括像素、正弦条纹、自然属性以及不可感知噪声等[17]，这些后门触发器的选用使后门攻击更加隐蔽。此外，后门攻击还可以通过预训练[18]或知识蒸馏[19]等方法强化后门触发器与 DNN 中间神经元之间的联系，使这些神经元在预测阶段对后门触发器具有强激活效果，进一步提升后门攻击的隐蔽性和危害性。

后门攻击在集中式学习和联邦学习场景下的实现方式有所区别。首先，在集中式学习场景中，后门攻击通常是通过数据投毒来完成的[20]。例如，在 CIFAR-10 数据集的“汽车”和“飞机”分类任务中，攻击方可以将训练数据中的所有绿色汽车标记为“飞机”，并试图使模型在预测阶段将绿色汽车分类为“飞机”。而在联邦学习场景中，由于训练数据分散在各个参与方本地，攻击方难以访问所有的本地数据。因此，联邦学习中的后门攻击通常以模型投毒的方式进行[21-24]。也就是说，攻击方在本地训练阶段将后门触发器注入本地模型中，构建出包含恶意后门的本地更新，当它与其他参与方的本地更新聚合后，联邦学习全局模型就会展现出后门攻击的固有特性。

针对现有的后门攻击方法，主要的防御思路包括 2 种：后门检测和后门消除。其中，基于后门检测的防御方法旨在识别目标模型中是否存在后门触发器[25-26]，或从训练数据中直接过滤可疑样本进行重新训练[27]。然而，这些被动式后门检测的防御方法仅能判断出模型是否存在后门攻击，无法消除后门攻击给目标模型所带来的负面影响。因此，研究者们开始探索如何通过后门消除的方式来净化后门模型。目前基于后门消除的防御方法主要通过在一部分干净数据上进行模型微调[28]，以及采用模型修剪等方法减少微调过程中可能带来的过拟合现象。除此之外，还有一些如数据增强、正则化、模型修复等方法[29]也被陆续提出以减轻后门攻击的效果。但是，基于后门消除的防御方法会降低主任务的分类精度，且算法效率问题难以解决[30-32]。

因此，针对上述问题，本文提出了一种基于对比训练的联邦学习后门防御方法 ContraFL，该方法能够在有效防御联邦学习后门攻击的同时，保持较高的全局模型分类精度。具体来说，ContraFL 首先在服务器执行触发器生成算法，构造生成器池，以还原全局模型中可能存在的后门触发器。进一步地，服务器将触发器生成器池下发给各参与方，各参与方将生成的后门触发器添加至本地样本，以实现数据增强，最终通过对比训练有效消除后门攻击的负面影响，提升模型的鲁棒性。本文的主要贡献包括以下 3 个方面。

1) 设计了一种简单有效的联邦学习后门防御策略，即通过对比训练来破坏特征空间中后门样本的聚类，使联邦学习全局模型分类结果与后门触发器特征无关，进而增强全局模型对后门攻击的鲁棒性。

2) 提出了一种基于对比训练的联邦学习后门防御方法 ContraFL，首先采用最大熵阶梯近似器[33]进行触发器重建，以实现后门数据增强，进而利用本地数据进行对比训练，以消除全局模型中潜在的后门属性的影响。特别地，ContraFL 能够在本地训练过程中实现对嵌入后门的有效消除。

3) 选取 3 种典型的联邦学习后门攻击方法，在 4 个基准数据集上进行实验测试，并与 4 种先进的联邦学习后门攻击防御方法进行对比，实验结果表明，ContraFL 能够有效防御后门攻击并提升模型鲁棒性，且优于大多数现有的防御方法。

1 相关技术

1.1 联邦学习

图 1 为 Google 于 2017 年提出的标准联邦学习架构[2]，其设计目标是在不同参与方本地数据上训练一个联合的机器学习模型，同时保证各参与方数据的隐私性。为了实现这一目标，联邦学习允许各个参与方(用户)下载全局模型到本地，并利用本地数据对模型进行训练以更新参数，最终将这些来自不同用户训练数据的更新参数汇总到服务器进行聚合平均，生成新一轮的全局模型[3-4]。

具体来说，在联邦学习训练开始之前，服务器首先根据参与方的服务需求初始化一个全局模型 $\tilde{w}_0$，并随机选择本轮参与方集合 s_t，将该模型下发至各个参与方；在接收到服务器发来的初始化模型后，各个参与方用自己的本地数据对模型进行训练，更新本地模型参数 w_k^t 并上传至服务器，其中，t 表示联邦学习的训练轮次，$k \in s_t$ 表示第 k 个参与方；在接收到来自参与方的本地模型参数后，服务器采用聚合平均算法 FedAvg[2]对所有参数进行加权平均以更新生成下一轮的全局模型。FedAvg 算法的聚合规则为

$$\tilde{w}^{t+1} = \sum_{k \in s_t} \frac{D_k}{D_{s_t}} w_k^t \tag{1}$$

其中，D_k 表示参与方的数据规模，D_{s_t} 表示本轮所有参与训练的数据的总规模。上述过程被不断迭代执行，直到服务器初始化的全局模型趋于收敛。

1.2 对比训练

在现有的机器学习方法中，监督学习方法通过大规模标记数据训练展现卓越的性能，占据主导地位。然而，在现实场景中，出于隐私安全、标注代价等原因，大量高质量的标签往往难以获得。因此，作为一种不依赖或仅依赖少量标签的训练范式，自监督学习得以蓬勃发展。对比学习作为自监督学习的一种，凭借其优越的性能，被广泛研究与应用。对比学习的核心思想是通过构造相似实例（正样本对）和不相似实例（负样本对），训练得到一个模型，使相似实例在特征空间（投影空间）中足够接近，而不相似实例在特征空间中相距足够远，以提高深度学习的特征表达能力[34]。

对比训练的具体过程如图 2 所示。首先，通过对数据 x_1、x_2 和 x_3 进行数据增强得到 x_1'、x_2' 和 x_3'。

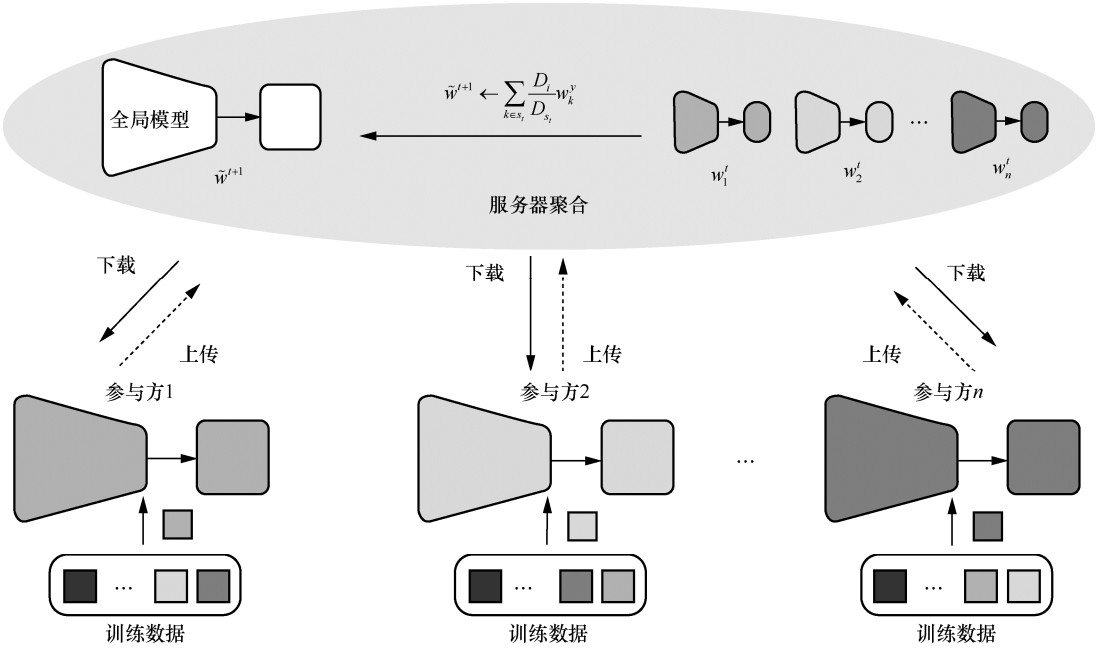


图 1 联邦学习架构

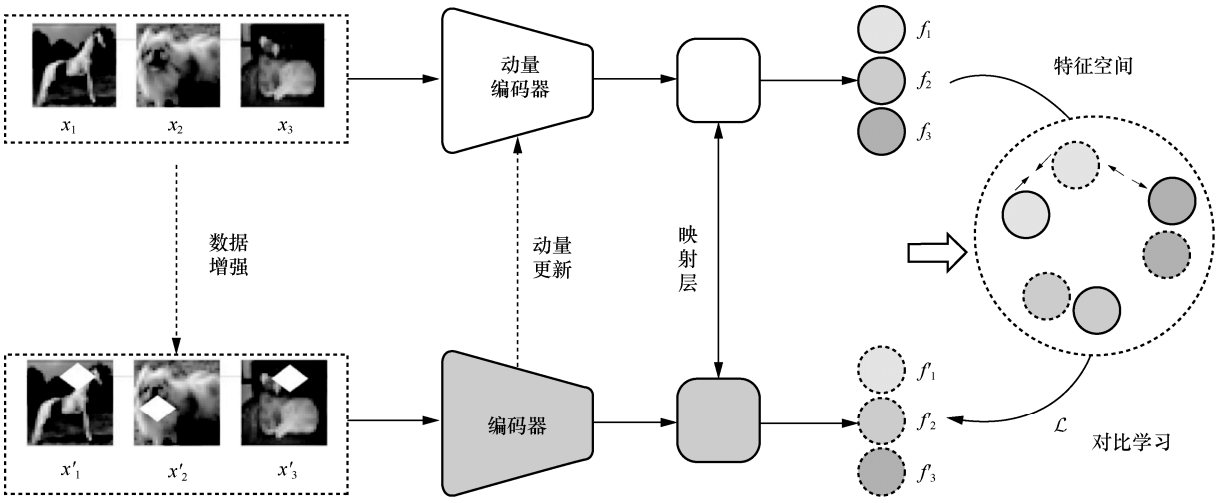


图 2　对比训练的具体过程

其中，原数据与其对应的增强数据互为正样本对，与其他数据互为负样本对。将原数据和增强数据分别输入 2 个编码器，再经过映射层得到各种特征。值得注意的是，映射层一般进行两次非线性变化，用以过滤部分特征。该设置在 SimCLR[35]模型中首次应用，能显著提升对比训练效果。最终，计算对比训练损失，更新编码器。为了保持特征的一致性，大多工作参照 MOCO（momentum contrastive）[36]模型中的设置，采用动量更新的方式更新另一编码器。近年来，另一些工作（如 BYOL（bootstrap your own latent）[37]和 SimSiam[38]）允许模型绕过负样本对，仅通过构造正样本对并采用梯度停止运算来避免琐碎的操作。也有学者提出在联邦学习中应用对比学习的工作[39-40]，通过将现有的对比训练框架直接迁移到联邦学习场景或者采用模型层面的对比框架，利用对比训练强大的表征学习能力，解决联邦学习中数据异构问题，但未考虑到后门攻击等安全隐患。本文应用对比学习进行鲁棒训练，以增强模型对后门攻击的防御能力。

2　ContraFL 方法

为了实现对联邦学习后门攻击的有效防御，本文在服务器训练后门触发器生成器，并在本地训练时在正常样本上添加后门触发器作为数据增强的一种方式进行对比训练，以提高模型对后门特征的鲁棒性。本节首先阐述联邦学习架构下的后门攻击模型，然后给出基于对比训练的联邦学习后门防御方法。本文的符号说明如表 1 所示。

表 1　符号说明

符号	描述
w	模型参数
$t \in T$	联邦学习通信轮次
$k \in s_t$	联邦学习参与方
E	联邦学习本地训练轮次
D	训练数据集规模
r	被污染数据占总数据的比例
Δ	触发器分布
$F(\cdot)$	全局模型
β_i	后门还原阈值
G_i	触发器生成器
T_i	互信息估计器
$\delta \sim \mathrm{N}(0,1)$	随机噪声
τ	后门攻击成功率阈值
$P_w(\cdot)$	编码器输出
$C_w(\cdot)$	分类层第一层的输出
A_x	后门数据增强
Z'	目标网络编码器输出
$\bar{Z}$	模型正则化输出

2.1　威胁模型和防御目标

1) 威胁模型

给定干净的本地样本 D_k，攻击方旨在通过添加特定触发器 Δ 来污染其中的一部分样本 r，并将受污染的原始标签修改为目标标签 y_t，以误导模型 F 输出错误分类[25,33]。同时，保证后门模型可以在干净的样本上正常表现。因此，攻击方通过最小化以下目标函数来篡改训练过程

$$\text{Loss} = \sum_{(x,y)\in D_k} \begin{cases} \ell(F(x+\Delta), y_t), r \\ \ell(F(x), y), 1-r \end{cases} \tag{2}$$

联邦学习后门攻击模型如图 3 所示。在联邦学习场景中，后门攻击具备更强的破坏性。首先，攻击方伪装良性参与方，通过在正常本地数据集上添加后门触发器，并赋予其指定的标签，以构造后门数据集。随后，攻击方在该数据集上进行训练得到后门模型，当该模型参与联邦学习服务器聚合后，全局模型也将存在后门模式，从而对带有后门触发器的输入产生误分类。

具体而言，本文假设攻击方具备以下能力：①攻击方拥有一定数量的干净数据集用于参与联邦学习任务；②攻击方可以操纵本地训练过程，能够将触发器植入良性样本中并为其分配所需的标签；③攻击方可以上传精心制作的参数，参与联邦学习聚合。但是，攻击方上传的模型在结构上必须与全局模型保持一致。此外，攻击方不能干涉服务器执行的操作。

2) 防御目标

为了更符合真实应用场景，参考先前的工作[25,33]，本文定义防御方存在以下限制。

①联邦学习的参与方仅能获得每轮服务器下发的聚合后的模型（全局模型），但并无其中蕴含的后门模式的相关信息。

②服务器仅拥有少量的干净数据集（测试集），而无法接触到任何本地训练集。

值得注意的是，本文假设的干净数据集一般采集自互联网中的公开数据集，或从各个参与方的验证集中采集一些不敏感的数据[25]，在实际应用中是切实可行的。最终，本文定义防御方的目标是通过部署 ContraFL，在显著降低后门攻击的成功率的同时，保持模型在正常样本上的准确率。

2.2 ContraFL 框架

图 4 展示了本文提出的 ContraFL 框架，主要步骤如下：①各参与方利用本地数据进行对比训练，将训练完成后的模型上传至服务器；②服务器首先通过 FedAvg 算法对接收的模型进行聚合，得到下一轮全局模型；③服务器利用少量干净模型，以全局模型为判别器，训练触发器生成器池；④服务器将全局模型和触发器生成器池下发至各参与方，进行下一轮的联邦学习训练。

1) 触发器还原

由式(2)可知，当同样的触发器被添加至正常样本上时，触发器特征与指定标签之间的强联系会扭曲模型正常的决策边界，从而实现攻击方设定的分类。基于上述事实，当获得一个后门模型后，本文的目的是

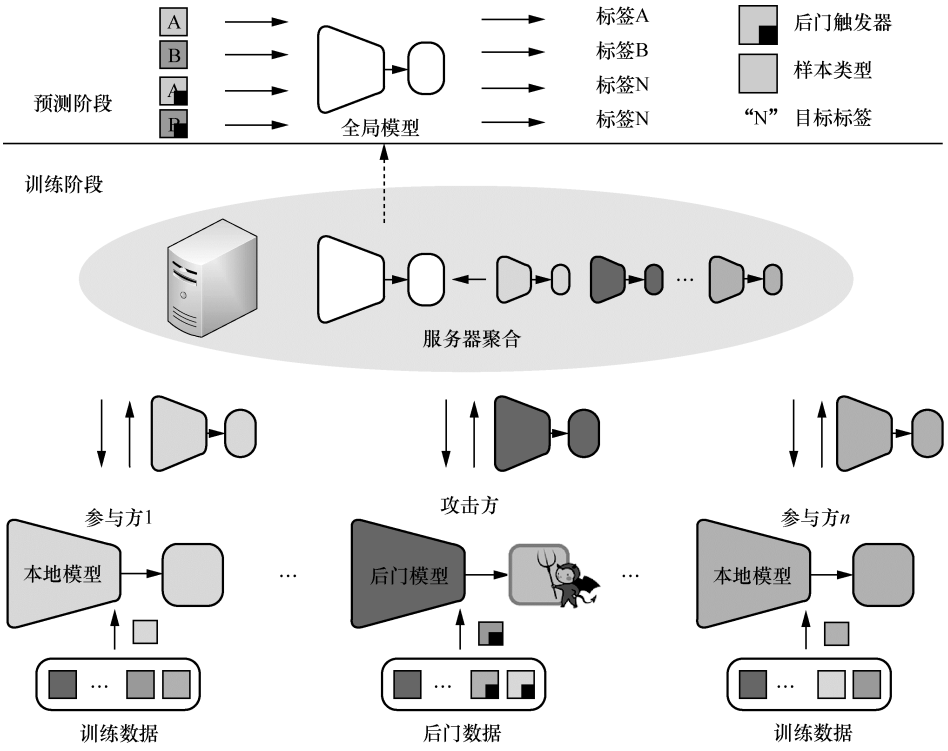


图 3 联邦学习后门攻击模型

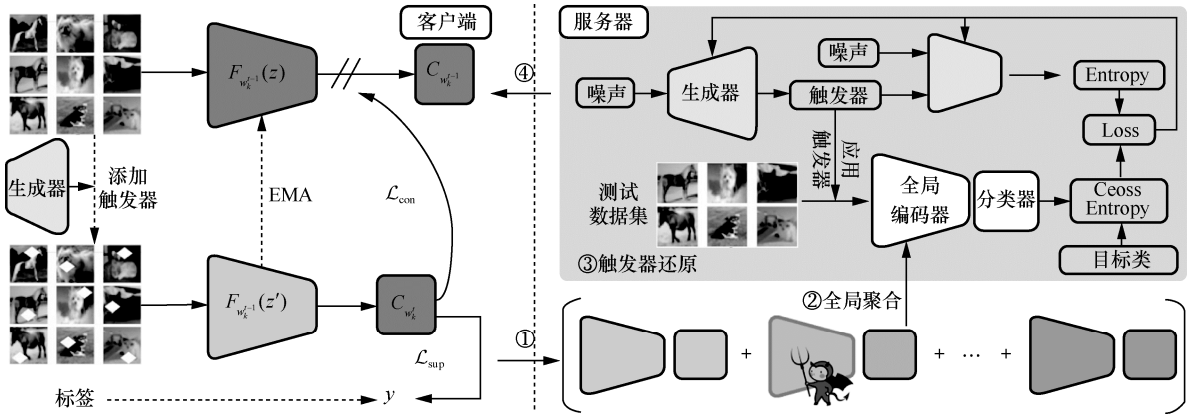


图 4　ContraFL 框架

寻找一个触发器分布 Δ，使 $F(x)=y$ 在添加 Δ 后转换为 $F(x+\Delta)=y_t$，其中，y 为该样本的真实标签，y_t 为攻击方指定的标签。自然地，可利用对抗生成网络（GAN）来实现该目标，通过部署一个生成模型来生成未知触发器分布 Δ，并采用后门模型作为辨别器以区分生成的触发器分布 Δ 是否有效。

然而，传统的对抗生成网络已被证明在估计高维触发模式和微分熵时会出现模型下降问题，难以适用本文的场景。因此，本文通过引入了最大熵阶梯近似器（MESA）[32]来克服这一问题。MESA 通过集成一组子模型 $G=\{G_1,G_2,\cdots,G_n\}$ 来近似未知的触发器分布 Δ。其中，每个子模型 G_i 学习触发器的一部分 $\Delta_i=\{\gamma,F_{\tilde{w}^{t+1}}(x+\gamma)\geqslant\beta_i\}$，$F_{\tilde{w}^{t+1}}$ 为后门模型。具体而言，本文首先从 [0,1] 中均匀地选择 n 个阈值 $\beta=\{\beta_1,\beta_2,\cdots,\beta_n\}$，并根据每一个阈值 β_i 初始化一个生成器 G_i 和一个互信息估计模型 T_i [38]。然后生成随机噪声 $\delta\sim$N(0,1) 和 $\delta'\sim$N(0,1)，并将 δ 输入生成器 G_i 得到部分生成触发器 $G_i(\delta)$。进一步，将生成的触发器添加至 x，再输入具有后门属性的全局模型，得到预测结果 $F_{\tilde{w}^{t+1}}(x+G_i(\delta))$。最后，基于 ℓ_R 优化生成器 G_i

$$\ell_R=\frac{1}{b}\sum_{x\in D_c}(\max(0,\beta_i-F_{\tilde{w}^{t+1}}(x+G_i(\delta))[y_t])-\eta T_i(G_i(\delta);\delta')) \tag{3}$$

当生成器收敛后，本文选择后门攻击成功率高于设定的阈值 τ 的生成器加入生成器池。在完成训练后，服务器将训练得到的触发器生成器下发给各个参与方。此外，为了避免服务器与本地参与方之间的通信时延，后门还原与本地训练采用并行方式，即服务器先完成全局模型聚合，将聚合后的全局模型下发至本地参与方后，再进行触发器还原，以保证联邦学习进程能够持续进行。

2) 对比训练

本地训练阶段，客户端初始化一对孪生网络，即在线网络与目标网络。该孪生网络的结构相同，皆由编码器与分类器组成。其中，分类器由两层全连接层构成。第一层的输出维度与编码器的输出维度相同，用于对比训练时输出在线网络的回归目标以进行对比；第二层的输出维度与具体的分类任务有关，在监督训练时用以实现对输入标签的预测。在联邦学习的场景中（首轮除外），本文使用全局模型 $\tilde{w}^t$ 作为该轮本地的在线网络，而采用上一轮该用户训练的本地模型 w_k^{t-1}，即上一轮的在线网络作为该轮的目标网络。该设置有以下 2 个优势。首先，在第一轮训练中，良性客户端将服务器下发的全局模型作为在线网络和目标网络，并用本地数据进行对比训练，在客户端本身可靠的前提下，其训练的本地模型也被认为是可靠的。在接下来的联邦通信中，如前文所述，该模型会作为目标网络与全局模型一起进行对比训练，可有效缓解聚合入全局模型中的后门的负面影响。此外，用上一轮的本地模型作为目标网络，能够保持联邦学习迭代过程中的特征一致性，从而缓解模型漂移。

定义参与方 k 拥有本地数据集 D_k，当接收服务器下发的全局模型和触发器生成器池 G 时，利用训练完成的生成器 G_i 生成后门触发器 $\Delta=G_i(z)$。在训练过程中，从本地数据集 D_k 中随机采样 $(x,y)((x,y)\in D_k)$，并对 x 进行数据增强得到 $A_x=x+\Delta$ 和 $A'(x)$，其中，$A'(x)$ 为一种数据增强策略。进一步地，以对 x 数据增强后的 $A(x)$ 和 $A'(x)$ 作为输入，全局模型 $\tilde{w}^t$ 作为在线网络，上一轮本地模型 w_k^{t-1} 作为目标网络，进行对比

训练。具体而言，定义 $F_w(\cdot)$ 为以 w 为参数的完整模型。其中，编码器的输出为 $P_w(\cdot)$，分类层第一层的输出为 $C_w(\cdot)$，第二层的输出为完整模型的输出 $F_w(\cdot)$。将 $A(x)$ 和 $A'(x)$ 分别输入在线网络和本地网络，计算在线网络分类层输出 $Z_1 = C_{\tilde{w}^t}(A(x))$ 和目标网络编码器输出 $Z_1' = P_{w_k^{t-1}}(A'(x))$，再分别对 Z_1 和 Z_1' 进行 L_2 正则化，并用均方误差来衡量正则化后的两者误差，定义其损失为

$$\ell_1 = \| \bar{Z}_1 - \bar{Z}_1' \|_2^2 = 2 - 2\frac{<\bar{Z}_1, \bar{Z}_1'>}{\| \bar{Z}_1 \|_2 \| \bar{Z}_1' \|_2} \tag{4}$$

随后，进行对称操作，即将 $A'(x)$ 和 $A(x)$ 分别输入在线网络和本地网络，计算在线网络分类层输出 $Z_2 = C_{\tilde{w}^t}(A'(x))$ 和目标网络编码器输出 $Z_2' = P_{w_k^{t-1}}(A(x))$，分别对 Z_2 和 Z_2' 进行 L_2 正则化得到 $\bar{Z}_2$ 与 $\bar{Z}_2'$，以相同的方式计算损失 ℓ_2。最终，对比学习的损失可定义为

$$\ell_{\text{contrastive}} = \ell_1 + \ell_2 \tag{5}$$

此外，为了充分利用联邦学习各参与方的本地标签信息，提升联邦学习主任务的分类准确率，本文计算了有监督损失为

$$\ell_{\text{supervised}} = \text{CrossEntropyLoss}(F_{w_k^t}(A'(x)), y) \tag{6}$$

因此，对比训练的总损失为

$$\ell_{\text{total}} = \ell_{\text{contrastive}} + \ell_{\text{supervised}} \tag{7}$$

采用随机梯度下降法，通过 ℓ_{total} 对在线网络进行优化更新，并通过动量更新方式对目标网络进行缓慢更新，具体更新方式如下

$$w_k^{t-1} = \lambda w_k^{t-1} + (1-\lambda) w^t \tag{8}$$

在完成规定轮次的对比训练后，参与方将更新后的在线网络作为本轮的本地训练模型上传至服务器进行聚合。ContraFL 算法如算法 1 所示。

算法 1 ContraFL 算法

输入 w_0，z，D_c，τ

输出 $\tilde{w}^T$

//服务器执行

① for 每个通信轮次 $t \in [1,2,\cdots,T]$ do
② 随机选取参与方子集 S_t
③ for 每个参与方 $k \in S_t$ do
④ $w_k^t \leftarrow \text{LocalTraining}(k, \tilde{w}^t, G)$
⑤ $\tilde{w}^{t+1} = \sum_{k \in s_t} \frac{D_k}{D_{s_t}} w_k^t$
⑥ end for
⑦ 初始化 G，$\beta = \{\beta_1, \beta_2, \cdots, \beta_n\} \in [0,1]$
⑧ for $y_t \in L$ //触发器还原
⑨ 为每个 $\beta_i \in \beta$ 初始化对应的 G_i，T_i
⑩ while not converged do
⑪ 从 D_c 中随机采样批大小为 b 的 D'
⑫ 生成 $\delta \sim$N(0,1)，$\delta' \sim$N(0,1)
⑬ 根据式(3)计算损失 ℓ_R
⑭ 通过文献[38]更新 T_i，通过 SDG 更新 G_i
⑮ if ASR of $G_i > \tau$
⑯ $G \leftarrow G \bigcup G_i$
⑰ end if
⑱ end while
⑲ end for
⑳ end for
㉑ return $\tilde{w}^T$

//客户端执行:

㉒ LocalTraining$(k, \tilde{w}^t, G)$;//本地训练
㉓ 用 $\tilde{w}^t$ 更新在线网络
㉔ 用 w_k^{t-1} 更新目标网络
㉕ for each local epoch $k \in [1,2,\cdots,E]$ do
㉖ for each batch $b = \{x, y\} \in D_k$ do
㉗ 从 G 中选择触发器生成器生成 Δ
㉘ 后门数据增强 $A_x = x + \Delta$
㉙ 随机数据增强 $A'(x) = x + z$
㉚ $w_k^t = w_k^t - \eta \nabla \ell_{\text{total}}(A(x), A'(x), y)$
㉛ $w_k^{t-1} = \lambda w_k^{t-1} + (1-\lambda) w^t$
㉜ end for
㉝ end for
㉞ return w_k^t

3 实验与结果分析

为了评估本文方法的有效性，本文在 PyTorch (1.8.0 + cu111)上实现了本文的 ContraFL 框架。其硬件为 Intel i9-9700K CPU、16 GB 内存和 NVIDIA GeForce GTX2080 GPU，软件为 Window10 操作系统和 Pycharm。

3.1 数据集与实验设置

3.1.1 数据集介绍

本文在常用的 4 个基线数据集上进行实验，分

别是 MINST 数据集、Fashion-MNIST（F-MNIST）数据集、CIFAR-10 数据集和 Extra-MNIST（EMNIST）数据集。值得注意的是，EMNIST 是对 MNIST 数据集的扩展，是一种更具挑战性的基线数据集，在实验中，该数据集主要用以验证非独立同分布（Non-IID）场景下 ContraFL 的性能。

3.1.2 实验相关设置

1) 模型设定

本实验中，模型由编码器与分类器组成。所有数据上的任务均采用 ResNet-18 作为编码器，并使用两层多层感知机（MLP）层作为分类头，第一层的维度与编码器输出维度相同，第二层的维度与具体的分类任务相关（10 分类）。关于服务器的生成器模型，本文采用与文献[33]中相同的结构。

2) 训练配置

关于实验中的联邦学习算法，本文采用 2017 年 Google 提出的标准联邦学习架构[2]。联邦学习训练协议在 100 个参与方中运行，在每轮的联邦通信中，随机选择 10 个参与方进行聚合。值得注意的是，本文设置每轮中固定有一定比例的恶意参与方被选中，这使联邦学习由于选择随机性产生的对后门攻击的抵制作用消失。在本地训练阶段，为了提高后门攻击的有效性，恶意参与方训练轮次和学习步长分别为 20 和 0.1，而良性参与方的训练轮次和学习步长分别为 10 和 0.05。上述设置仅针对基于不同触发器类型与大小的后门攻击，关于实验中所实现的基线后门攻击与防御方法，本文参照其原有设置。此外，除了关于干净数据集的比率的消融实验外，其他实验中 ContraFL 的实现均假设服务器可接触 3%的干净数据集。本文从验证集中抽取 3%作为服务器可接触的干净数据集。同时，本文验证了不同比例的干净数据集对 ContraFL 的影响。

3) 评估指标

为了直观地体现本文所提 ContraFL 的有效性，实验定义 2 种评估指标，分别为：后门任务准确率（ASR），表示攻击方添加触发器的后门样本中被分类为攻击方所选择标签的样本数量与攻击方后门样本总数量之间的比值；主任务准确率（MA），代表非攻击方目标样本被分类为正确原始标签的样本数量与除攻击方目标样本以外的总样本数量之间的比值。一个鲁棒的联邦学习框架能够在保持高 MA 的同时显著降低 ASR。

4) 基线方法

本文验证了 ContraFL 对 3 种最新且有效的后门攻击的防御效果，即 DBA[22]、PGD[21]和 WANET[17]。其中，DBA 是一种分布式后门攻击，通过多恶意参与方嵌入不同的后门触发器从而实现更加隐蔽且持久的攻击。PGD 是一种边缘后门攻击，迫使模型对一些简单且位于输入分布尾部的输入进行错误分类。WANET 是一种不可见后门攻击，相较于其他后门攻击，更加难以检测。相应地，本文还选择了 4 种典型的后门防御方法作为本文的比较基线，即 FoolsGold[27]、CONTRA[14]、Baffle[26]和 RLR[31]。其中前 2 种属于基于后门检测的防御方法，后 2 种属于基于后门消除的防御方法。上述攻击与防御方法皆按照其原设置实现。此外，本文验证了 ContraFL 对基于不同触发器类型与触发器大小的后门攻击的鲁棒性。

3.2 与基线防御方法的比较

表 2 显示了 ContraFL 与基线防御方法的性能对

表 2　ContraFL 与基线防御方法的性能对比

数据集	攻击方法	无防御场景		FoolsGold[27]		CONTRA[14]		Baffle[26]		RLR[31]		ContraFL	
		ASR	MA	ASR	MA	ASR	MA	ASR	MA	ASR	MA	ASR	MA
MNIST	DBA	100%	90.43%	97.62%	79.32%	8.05%	87.81%	8.95%	83.66%	7.18%	85.87%	2.98%	90.61%
	PGD	92.47%	93.01%	12.45%	80.85%	5.46%	88.87%	7.14%	83.53%	2.13%	88.93%	1.32%	92.19%
	WANET	93.76%	90.59%	18.95%	81.66%	7.51%	84.82%	10.61%	79.3%	4.06%	85.49%	4.03%	89.15%
F-MNIST	DBA	100%	89.67%	98.31%	79.81%	6.67%	87.82%	10.02%	85.09%	5.51%	84.81%	2.88%	91.19%
	PGD	91.58%	92.56%	10.22%	81.61%	1.87%	87.11%	2.79%	84.99%	3.69%	86.77%	1.36%	90.23%
	WANET	92.85%	86.85%	19.05%	79.34%	8.55%	80.27%	7.42%	83.52%	3.36%	86.38%	1.35%	85.93%
CIFAR-10	DBA	94.12%	72.14%	91.89%	71.09%	8.93%	72.55%	10.09%	70.36%	8.82%	69.10%	3.44%	73.96%
	PGD	58.61%	83.56%	7.45%	81.53%	4.58%	81.24%	5.98%	80.05%	5.59%	80.71%	2.08%	81.44%
	WANET	92.19%	85.23%	17.61%	79.35%	6.91%	79.78%	9.78%	80.07%	6.08%	74.23%	2.60%	84.98%

比，所有方法都使用了与原论文相同的设置。实验表明，ContraFL 在 3 个基准数据集上可以将上述 4 种后门攻击的 ASR 降低到 4%左右。这超过了其他 3 种防御方法的性能。总体来说，实验中的所有防御方法对集中式的后门攻击都很有效，而 FoolsGold 却无法防御以分布式进行的 DBA。具体而言，作为基于后门检测的防御方法，CONTRA 和 Baffle 在防御后门攻击方面具有相似的水平，而 FoolsGold 的效果则略逊一筹。作为基于后门消除的防御方法，RLR 通过鲁棒学习率训练，以缓解后门，其展现出更好的防御效果。然而，所有的基线防御方法都会导致模型在主任务上的表现有所退化，即 MA 有明显的下降。而 ContraFL 在实现更优的防御效果的同时，将模型准确率的下降控制在 2%左右，甚至在防御 DBA 时，MA 对比“Before”有所提升。据本文分析，这是由于相较于其他类型的后门攻击，DBA 由多个恶意参与方共同发起，其本身对模型造成了相对严重的性能退化。因此，当该攻击被有效抵制后，主任务准确率有了显著的回升。此外，实验结果也表明了不管触发器类型是分布式的（DBA）还是不可见的（WANET），ContraFL 仍然有效。

3.3 ContraFL 鲁棒性分析

为了进一步探究 ContraFL 的鲁棒性，本文通过 t-SNE 可视化了在面对后门攻击时，部署 ContraFL 和未部署 ContraFL 全局编码器在特征空间中的分布。t-SNE 是一种降维技术，主要用于在低维空间中表示高维数据集，相较于其他的降维算法，对数据具有更好的可视化效果。更具体地说，本文修改了 CIFAR-10 数据中 10%的测试数据，即 10 000 张图像中 1 000 张被添加了后门触发器，其中每一类修改的图像数量是相同的。随后，将后门测试数据用于 t-SNE 可视化。图 5 展示了部署与未部署 ContraFL 全局编码器的特征可视化结果。从图 5(a)中可以看出，在未部署 ContraFL 的情况下，同一类别的良性样本的特征表示各自形成了单独的聚类，而后门样本则形成了一个新的聚类（黑色），这意味着对于一个后门模型来说，当正常样本被添加了攻击方设定的触发器作为输入后，该模型会以一个较高的置信度将其分类为攻击方指定的类。

如图 5(b)所示，在部署 ContraFL 后，由后门样本形成的聚类成功被破坏，绝大多数后门样本都与同一类别的良性样本重新聚合在一起。此外，良性样本以相对集中的方式定位在自己的聚类中，聚类内样本之间的距离非常接近。同时，聚类之间存在明显的距离。这说明在部署 ContraFL 后，模型中的后门成功被破坏，且模型在主要分类任务上表现仍然出色。

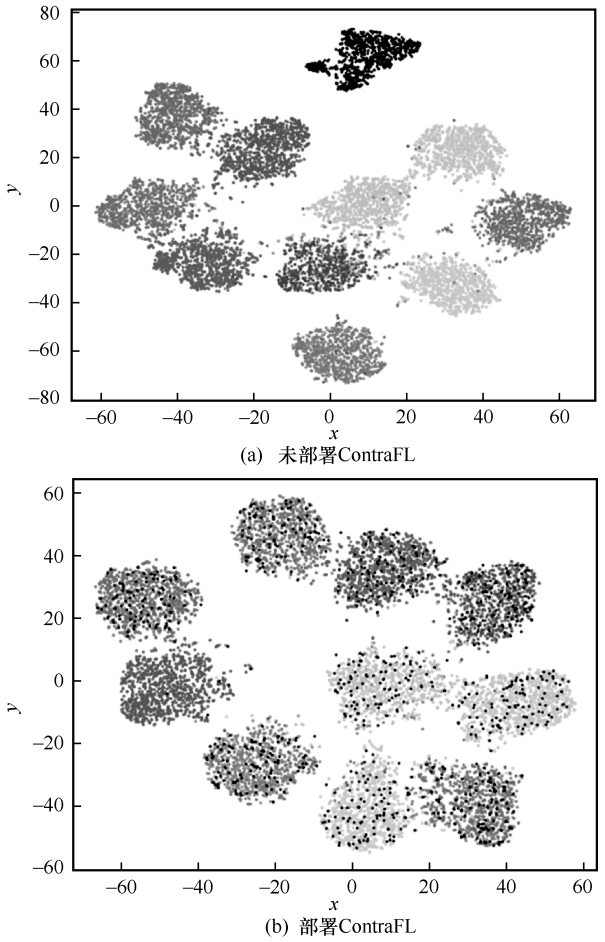


图 5 部署与未部署 ContraFL 全局编码器的特征可视化结果

为了进一步验证 ContraFL 对不同触发器类型与大小的鲁棒性，本文选择了 3 种不同类型的触发器（即像素块、水印和随机噪声），并将这些后门触发器以不同的尺寸嵌入。对于水印攻击，将数字“0”作为水印图案，并调整其水印系数在 0.1 到 0.5 之间变化，其跨度为 0.2。在基于像素块的后门攻击中，本文将触发器大小从 3×3 扩展到 5×5 和 7×7。对于随机噪声攻击，本文将噪声强度从 10 增强至 20 和 30。部分实验中使用的添加后门触发器的样本如图 6 所示。针对上述基于不同触发器类型与大小的后门攻击，本文在 CIFAR-10 数据集上与 4 种基线方法进行对比，结果如表 3 所示。总体来说，4 种基线方法和 ContraFL 都取得了可观的防御效果，但 ContraFL 呈现出更佳的性能。具体而言，从不同触发器类型来看，基于水印的后门攻击比另外

2 种类型的后门攻击的准确率更高。但触发器大小似乎并未对后门攻击的成功率和不同防御策略的性能造成明显的影响。

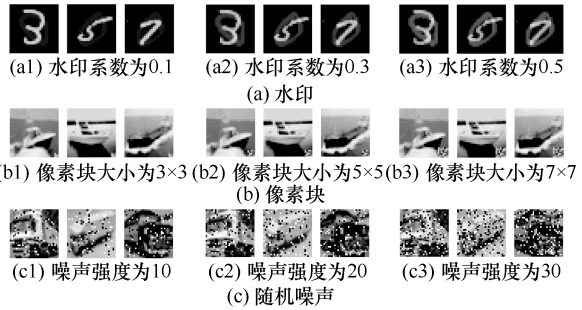


图 6　不同触发类型和大小的后门图像示例

进一步地，针对基于像素块的后门攻击，本文可视化了部署和未部署 ContraFL 的混淆矩阵（如图 7 所示），以更直观地展现 ContraFL 的性能。在 MNIST 任务中，原始标签为“9”的测试样本被嵌入 4×4 的像素块，并篡改其标签为“0”。对于 F-MNIST 和 CIFAR-10 任务，与 MNIST 数据集嵌入方式相同，分别为第 0 类（“T 恤”）分配第 6 类（“衬衫”）的标签，为第 5 类（“狗”）分配第 3 类（“猫”）的标签。结果表明，在未部署 ContraFL 场景下，MNIST 数据中大多数带有触发器的样本“9”被错误地分类为“0”。同样的情况也出现在 F-MNIST 和 CIFAR-10 数据集中。然而，在部署 ContraFL 后，绝大多后门输入被分类到其原始标签。

此外，考虑到 ContraFL 的实现需要依赖服务器存在一定比例的干净数据集，用以训练后门触发器生成器，本文验证了干净数据集比率（即未参与模型训练的干净样本的数据量与参与训练的样本总量的比值）对 ContraFL 性能的影响。图 8 展示了 ContraFL 在 3 个数据集上 MA 和 ASR 随防御方持有的干净数据集的不同比率（1%~10%）下的变化情况。直观地看，ContraFL 性能随着干净数据集比率的提升而提升，因为干净数据集的数量越多，后门触发器的还原效果越好，进而对比训练的效果也会有所提升。尽管如此，ContraFL 仅使用 1%的干净数据集，依然能够达到可观的防御效果，这在实际场景中是可行的。

3.4　适用性分析

1) 客户端数量

客户端的数量是联邦学习场景中的一个关键设置。为了验证本文方法的适用性，本节评估了在 MNIST、F-MNIST 和 CIFAR-10 数据集上不同客户端数量对 ContraFL 性能的影响。图 9 展示了 MA 和 ASR 在不同客户端数量下的变化。从图 9 中可以看出，MA 的值并未展现出明显的变化规律，而 ASR 随着客户端数量的增加有一定的下降趋势。据本文分析，这主要是因为随着客户端数量的增加，更多的本地模型参与聚合，在一定程度上降低了后门模型在全局模型中的权重，从而导致后门攻击的成功率降低。总体来说，ContraFL 的性能几乎不受客户端数量的影响，在现实应用场景中具备适用性。

2) Non-IID 数据分布

Non-IID 是联邦学习场景下常见的一类问题，其中各参与方所拥有的数据量与数据类别都不均

表 3　　**基于不同类型和大小触发器后门攻击下的 ContraFL 性能表现**

触发器类型	触发器大小	无防御场景		FoolsGold[27]		CONTRA[14]		Baffle[26]		RLR[31]		ContraFL	
		ASR	MA	ASR	MA	ASR	MA	ASR	MA	ASR	MA	ASR	MA
水印	0.1%	78.45%	84.36%	7.88%	79.72%	3.40%	82.96%	4.05%	81.89%	2.97%	80.63%	2.17%	83.43%
	0.3%	79.85%	83.49%	8.63%	79.26%	3.63%	81.28%	5.59%	81.04%	3.45%	81.22%	2.16%	83.78%
	0.5%	78.91%	84.31%	6.74%	79.94%	3.99%	82.18%	4.91%	80.75%	3.31%	81.93%	2.01%	82.35%
像素块	3×3	76.54%	85.47%	8.17%	79.46%	2.99%	82.50%	3.08%	79.46%	2.97%	79.13%	0.96%	84.42%
	5×5	76.22%	83.24%	7.74%	79.02%	3.24%	83.26%	4.15%	81.85%	3.21%	81.62%	1.15%	83.98%
	7×7	76.65%	84.33%	9.68%	78.9%	2.61%	82.20%	3.92%	80.82%	1.98%	80.72%	1.18%	83.47%
随机噪声	10%	77.54%	87.13%	7.18%	78.67%	3.5%	83.72%	4.76%	82.46%	2.27%	81.12%	1.32%	86.56%
	30%	76.90%	85.78%	8.61%	79.37%	4.25%	82.64%	6.72%	81.54%	3.42%	82.09%	1.92%	83.19%
	50%	76.19%	84.16%	7.57%	79.12%	4.74%	83.11%	5.68%	82.08%	4.40%	81.64%	1.45%	83.09%

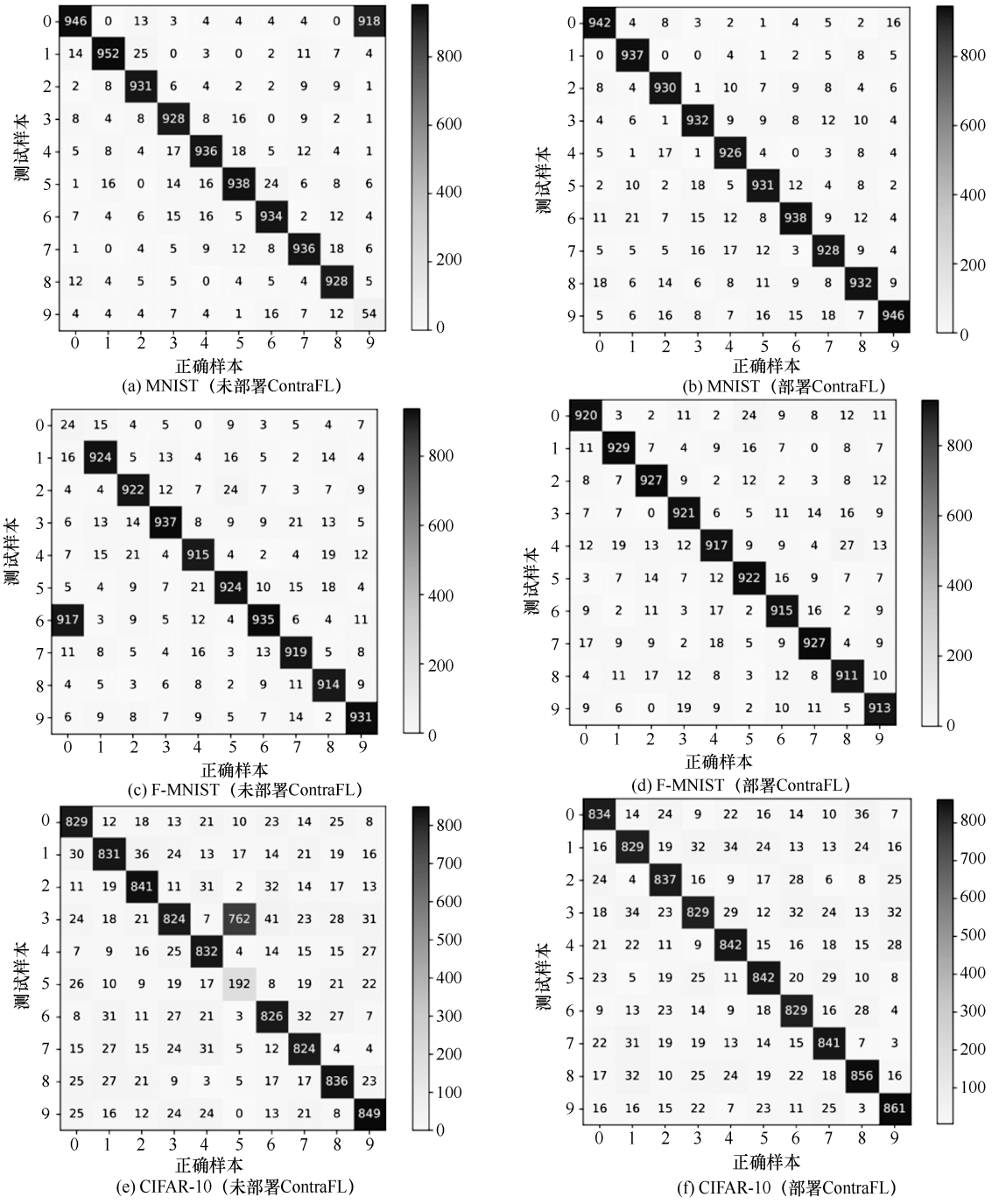
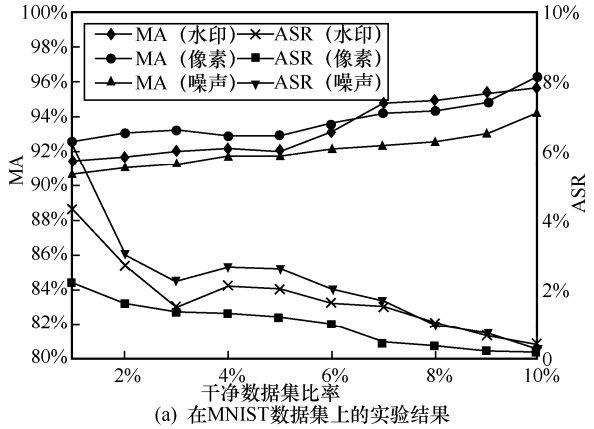


图 7　部署与未部署 ContraFL 方案在基线数据集上的混淆矩阵

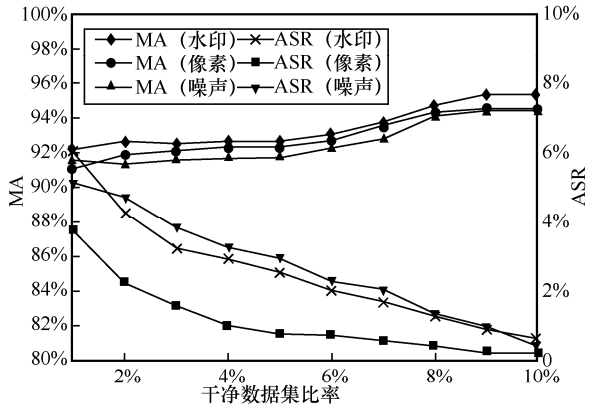
衡。根据现有技术[39]，本文使用狄利克雷分布 Dir(α) 对 Non-IID 数据分布进行建模，其中越小的 α 表示越高的数据异质性，即数据分布越不均衡。具体而言，本文在 EMNIST 数据集上针对 DBA 验证了 α 为 0.05、0.1 和 10 的防御效果，其中，α=0.05 和 α=0.1 表示 Non-IID；α=10 表示 IID。表 4 显示了 ContraFL 与 4 种基线后门防御方法的性能对比。结果表明，ContraFL 在 Non-IID 设置上显著优于所有基线方法，并将 DBA 的 ASR 降低至约 7%，而导致 MA 的退化可忽略不计。

随着数据异质性程度的增加，所有的 ASR 和 MA 都有所降低，但 ASR 仍高于 80%，说明尽管各个参与方所拥有的数据是不均衡的，后门攻击仍然生效。然而，除了 RLR 以外，其他 3 种基线防御方法几乎都不再起作用，主要原因在于 FoolsGold 和 CONTRA 都依赖于检测异常更新

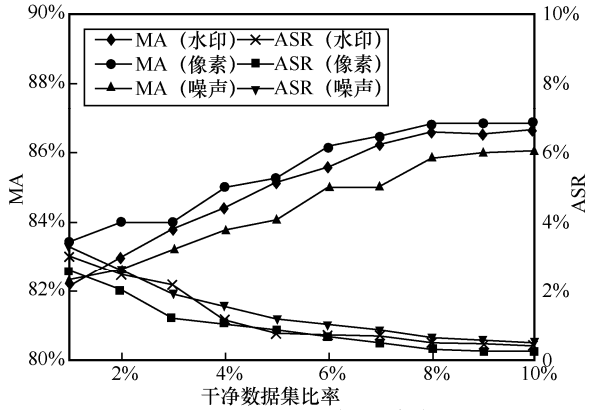
来抵御后门攻击，而在 Non-IID 设置下，由于各个参与方数据不均衡，训练所得的模型也存在差异，很难对异常更新和正常更新进行区分。Baffle 方法依据各参与方的本地更新动态调整每个客户端的权重，在该场景下也不适用。而 RLR 通过使用鲁棒的学习步长，几乎不受该设置影响。对于 ContraFL 而言，其主要依赖于本地对比训练来对后门进行防御，因此也不会受到 Non-IID 设置的影响，展现出更强的鲁棒性。



(a) 在MNIST数据集上的实验结果



(b) 在F-MNIST数据集上的实验结果



(c) 在CIFAR-10数据集上的实验结果

图 8　ContraFL 在 3 个数据集上 MA 和 ASR 随防御方持有的干净数据集比率的变化情况

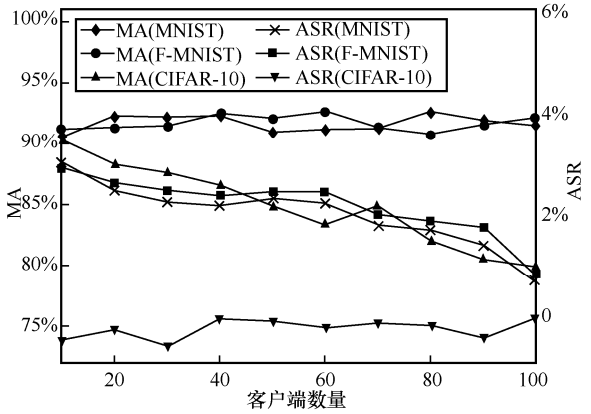


图 9　MA 和 ASR 在不同客户端数量下的变化

3) 本地训练轮次

本节还探讨了本地训练轮次对本文方法的影响。具体而言，本文在 CIFAR-10 数据集上针对 DBA[22]进行了防御验证，本地训练轮次从 10 增加至 50，步长为 10 轮。图 10(a)显示了随着本地训练轮次增加，MA 和 ASR 的变化趋势。从图 10(a)中可以看出，随着本地训练轮次的增加，ASR 逐渐下降，而 MA 快速增加，这意味着更多的本地训练轮次有助于增强防御效果和主任务准确率。然而，MA 的提升会随着轮次的增加而逐渐减缓。此外，更多的训练轮次意味着更多的计算开销，因此 ContraFL 尽可能选择更小的本地训练轮次以满足现实应用场景。值得注意的是，本文在本地训练轮次为 20 时已经取得了非常不错的后门防御效果。

由于对比训练过程中的批次大小对模型的性能往往具有较大影响，因此，本节进一步验证了本地训练批次对 ContraFL 的性能影响，其中预训练数据集选用 CIFAR-10。如图 10(b)所示，MA 随着批次大小的增加而有一定的提升，但并不显著，主要原因在于 ContraFL 采用的对比训练范式已被证明受批次大小影响较小。此外，本节在训练过程中增加了本地的监督信息，这进一步缓解了批次大小的影响。因此，本文方法符合联邦学习本地参与方的计算资源受限的现实场景。

4　结束语

本文提出了一个基于对比训练的新型联邦学习后门防御框架 ContraFL。具体来说，ContraFL 通过在服务器训练触发器生成器用以还原全局模型中可能存在的后门触发器。进一步地，将触发器生成器池下发给各参与方，各参与方生成后门触发器

表 4　　不同联邦学习后门防御方法在 EMNIST 数据集上的比较结果

α	无防御场景		FoolsGold[27]		CONTRA[14]		Baffle[26]		RLR[31]		ContraFL	
	ASR	MA	ASR	MA	ASR	MA	ASR	MA	ASR	MA	ASR	MA
0.05	80.28%	68.76%	77.32%	66.09%	44.26%	61.92%	54.92%	61.4%	7.62%	65.77%	6.62%	68.29%
0.1	84.08%	71.82%	58.20%	67.51%	34.21%	65.67%	53.71%	63.31%	6.44%	70.02%	4.45%	70.81%
10	93.58%	75.14%	92.76%	69.05%	9.43%	71.42%	11.01%	70.10%	9.12%	69.78%	3.04%	74.56%

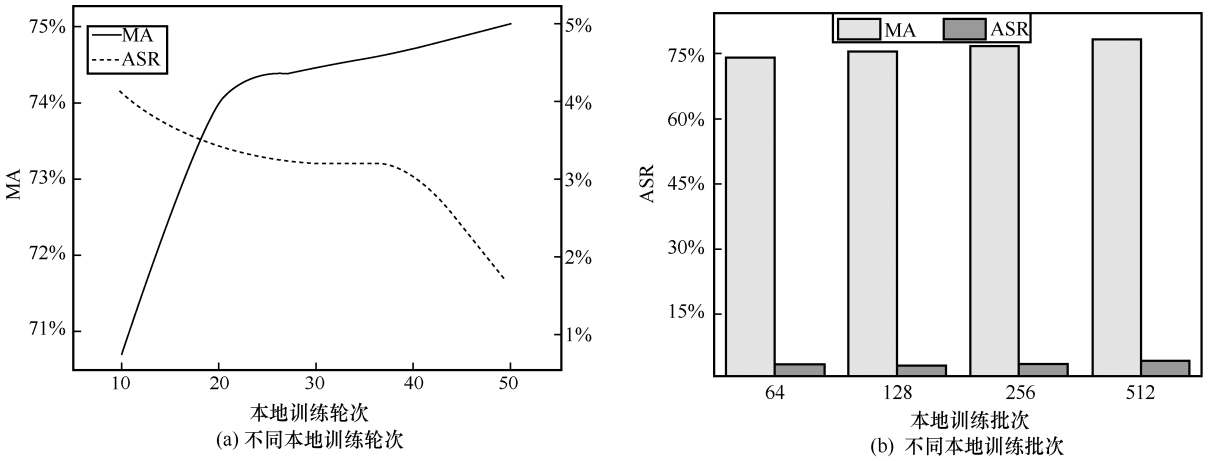


图 10　ContraFL 在不同本地训练轮次和批次下的性能表现

添加本地样本上用以实现数据增强，通过对比训练有效消除后门攻击的负面影响，提升模型的鲁棒性。大量的实验表明，ContraFL 能够有效防御后门攻击并提升模型鲁棒性，且优于现有的防御方法。在未来的工作中，笔者将进一步探索不依赖于干净样本的后门防御方法。

参考文献：

[1] 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 Press, 2017: 1175-1191.

[2] 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. New York: PMLR, 2017: 1273-1282.

[3] YANG Q, LIU Y, CHEN T J, et al. Federated machine learning: concept and applications[J]. ACM Transactions on Intelligent Systems and Technology, 2019, 10(2): 1-9.

[4] BONAWITZ K, EICHNER H, GRIESKAMP W, et al. Towards federated learning at scale: system design[J]. arXiv Preprint, arXiv: 1902.01046, 2019.

[5] LIU Y, FAN T, CHEN T, et al. FATE: an industrial grade platform for collaborative learning with data protection[J]. Journal of Machine Learning Research, 2021, 22(226): 1-6.

[6] LIM W Y B, LUONG N C, HOANG D T, et al. Federated learning in mobile edge networks: a comprehensive survey[J]. IEEE Communications Surveys & Tutorials, 2020, 22(3): 2031-2063.

[7] YIN X F, ZHU Y M, HU J K. A comprehensive survey of privacy-preserving federated learning: a taxonomy, review, and future directions[J]. ACM Computing Surveys, 2021, 54(6): 1-36.

[8] 周纯毅, 陈大卫, 王尚, 等. 分布式深度学习隐私与安全攻击研究进展与挑战[J]. 计算机研究与发展, 2021, 58(5): 927-943.
ZHOU C Y, CHEN D W, WANG S, et al. Research and challenge of distributed deep learning privacy and security attack[J]. Journal of Computer Research and Development, 2021, 58(5): 927-943.

[9] 陈宇飞, 沈超, 王骞, 等. 人工智能系统安全与隐私风险[J]. 计算机研究与发展, 2019, 56(10): 2135-2150.
CHEN Y F, SHEN C, WANG Q, et al. Security and privacy risks in artificial intelligence systems[J]. Journal of Computer Research and Development, 2019, 56(10): 2135-2150.

[10] BHAGOJI A, CHAKRABORTY S, MITTAL P, et al. Analyzing federated learning through an adversarial lens[C]//Proceedings of the 36th International Conference on Machine Learning. New York: ACM Press, 2019: 634-643.

[11] SONG M K, WANG Z B, ZHANG Z F, et al. Analyzing user-level privacy attack against federated learning[J]. IEEE Journal on Selected Areas in Communications, 2020, 38(10): 2430-2444.

[12] BAGDASARYAN E, VEIT A, HUA Y, et al. How to backdoor federated learning[C]//Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics. New York: PMLR, 2020: 2938-2948.

[13] FANG M H, CAO X Y, JIA J Y, et al. Local model poisoning attacks to byzantine-robust federated learning[C]//Proceedings of the 29th USENIX Conference on Security Symposium. New York: ACM Press, 2020: 1623-1640.

[14] SANA A, LUO B, LI F. CONTRA: defending against poisoning attacks in federated learning[C]//Proceedings of the 26th European Symposium on Research in Computer Security. Berlin: Springer, 2021: 455-475.

[15] 纪守领, 杜天宇, 李进锋, 等. 机器学习模型安全与隐私研究综述[J]. 软件学报, 2021, 32(1): 41-67.
JI S L, DU T Y, LI J F, et al. Security and privacy of machine learning models: a survey[J]. Journal of Software, 2021, 32(1): 41-67.

[16] GU T Y, DOLAN-GAVITT B, GARG S. BadNets: identifying vulnerabilities in the machine learning model supply chain[J]. arXiv Preprint, arXiv: 1708.06733, 2017.

[17] NGUYEN A, TRAN A. WaNet: imperceptible warping-based backdoor attack[J]. arXiv Preprint, arXiv: 2102.10369, 2021.

[18] YAO Y S, LI H Y, ZHENG H T, et al. Latent backdoor attacks on deep neural networks[C]//Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2019: 2041-2055.

[19] LIU Y Q, MA S Q, AAFER Y, et al. Trojaning attack on neural networks[C]//Proceedings of the 2018 Network and Distributed System Security Symposium. Reston: Internet Society, 2018: 1-15.

[20] LIN J Y, XU L, LIU Y Q, et al. Composite backdoor attack for deep neural network by mixing existing benign features[C]//Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2020: 113-131.

[21] WANG H, SREENIVASAN K, RAJPUT S, et al. Attack of the tails: yes, you really can backdoor federated learning[C]//Proceedings of the 34th Annual Conference on Neural Information Processing Systems. Massachusetts: MIT Press, 2020: 16070-6084.

[22] XIE C, HUANG K, CHEN P Y, et al. DBA: distributed backdoor attacks against federated learning[C]//Proceedings of the 8th International Conference on Learning Representations. Vancouver: ICLR, 2020: 1-15.

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

[24] 陈大卫, 付安民, 周纯毅, 等. 基于生成式对抗网络的联邦学习后门攻击方案[J]. 计算机研究与发展, 2021, 58(11): 2364-2373.
CHEN D W, FU A M, ZHOU C Y, et al. Federated learning backdoor attack scheme based on generative adversarial network[J]. Journal of Computer Research and Development, 2021, 58(11): 2364-2373.

[25] WANG B L, YAO Y S, SHAN S, et al. Neural cleanse: identifying and mitigating backdoor attacks in neural networks[C]//Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2019: 707-723.

[26] ANDREINA S, MARSON G A, MÖLLERING H, et al. BaFFLe: backdoor detection via feedback-based federated learning[C]//Proceedings of the 2021 IEEE 41st International Conference on Distributed Computing Systems (ICDCS). Piscataway: IEEE Press, 2021: 852-863.

[27] FUNG C, YOON C J M, BESCHASTNIKH I. The limitations of federated learning in sybil settings[C]//Proceedings of the 23rd International Symposium on Research in Attacks, Intrusions and Defenses. Berlin: Springer, 2020: 301-316.

[28] LIU K, DOLAN-GAVITT B, GARG S. Fine-pruning: defending against backdooring attacks on deep neural networks[C]//International Symposium on Research in Attacks, Intrusions, and Defenses. Berlin: Springer, 2018: 273-294.

[29] TRUONG L, JONES C, HUTCHINSON B, et al. Systematic evaluation of backdoor data poisoning attacks on image classifiers[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE Press, 2020: 3422-3431.

[30] XIE C L, CHEN M H, CHEN P Y, et al. CRFL: certifiably robust federated learning against backdoor attacks[C]//Proceedings of the 38th International Conference on Machine Learning. New York: PMLR, 2021: 11372-11382.

[31] OZDAYI M S, KANTARCIOGLU M, GEL Y R. Defending against backdoors in federated learning with robust learning rate[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(10): 9268-9276.

[32] QIAO X, YANG Y, LI H. Defending Neural backdoors via generative distribution modeling[C]// Proceedings of the 33th Annual Conference on Neural Information Processing Systems. Massachusetts: MIT Press, 2019: 14004-14013.

[33] LIU Y, FAN M Y, CHEN C, et al. Backdoor defense with machine unlearning[C]//Proceedings of the IEEE Conference on Computer Communications. Piscataway: IEEE Press, 2022: 280-289.

[34] CHUANG C Y, ROBINSON J, LIN Y C, et al. Debiased contrastive learning[C]//Proceedings of the 34th Annual Conference on Neural Information Processing Systems. Massachusetts: MIT Press, 2020: 8765-8775.

[35] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations[C]//Proceedings of the 37th International Conference on Machine Learning. New York: ACM Press, 2020: 1597-1607.

[36] CHEN X L, FAN H Q, GIRSHICK R, et al. Improved baselines with momentum contrastive learning[J]. arXiv Preprint, arXiv: 2003.04297, 2020.

[37] GRILL J-B, STRUB F, ALTCHE F, et al. Bootstrap your own latent-a new approach to self-supervised learning[C]//Proceedings of the 34th Annual Conference on Neural Information Processing Systems. Massachusetts: MIT Press, 2020: 21271-21284.

[38] CHEN X L, HE K M. Exploring simple Siamese representation learning[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2021: 15745-15753.

[39] LI Q B, HE B S, SONG D. Model-contrastive federated learning[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer

Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2021: 10708-10717.

[40] TAN Y, LONG G, MA J, et al. Federated learning from pre-trained models: a contrastive learning approach[C]//Proceedings of the 36th Annual Conference on Neural Information Processing Systems. Massachusetts: MIT Press, 2022: 19332-19344.

[作者简介]

张佳乐（1994− ），男，安徽蚌埠人，博士，扬州大学副教授、硕士生导师，主要研究方向为人工智能安全和区块链安全。

朱诚诚（2000− ），男，安徽临泉人，扬州大学硕士生，主要研究方向为联邦学习安全与隐私保护。

成翔（1988− ），男，新疆乌鲁木齐人，博士，扬州大学讲师、硕士生导师，主要研究方向为网络安全态势感知和系统安全。

孙小兵（1985− ），男，江苏姜堰人，博士，扬州大学教授、博士生导师，主要研究方向为软件安全和系统安全。

陈兵（1970− ），男，江苏南通人，博士，南京航空航天大学教授、博士生导师，主要研究方向为人工智能安全与隐私保护、网络安全和无人系统安全。