一、论文选题依据（包括本课题国内外研究现状述评，研究的理论与实际意义，对科技、经济和社会发展的作用等）

1．背景及意义

随着大数据、云计算、物联网等技术的发展，数据产生、收集、存储和利用的速度和规模不断扩大，数据的价值日益凸显，伴随而来的是层出不穷的隐私威胁和信任危机[1]。为了防止敏感数据的泄露，各个企业或部门将数据储存在本地，导致部门之间的数据无法实现高效流通和共享，形成了“数据孤岛”问题[2]。“数据孤岛”的存在不仅导致数据资源的低效利用和价值流失，也成为了限制机器学习发展的主要瓶颈之一。

在此背景下，联邦学习（Federated Learning，FL）[3-4]作为一种新兴的分布式机器学习框架，旨在解决“数据孤岛”和隐私安全问题。经典的联邦学习框架包含服务器和客户端，它允许各个客户端在不上传隐私数据的情况下，在本地设备上训练局部模型，然后将局部模型参数上传到中央服务器进行聚合，从而得到一个全局模型，然后经过不断迭代直至全局模型收敛，做到“数据不动模型动” ，即保护了隐私数据，又能充分挖掘出数据的潜在价值。目前，联邦学习被广泛应用与医疗、金融、工业等各个领域。

然而，联邦学习的分布式结构也极易遭受到攻击，投毒攻击就是危害联邦学习鲁棒性的主要安全威胁之一，主要包括数据投毒[5]和模型投毒[6]。

数据投毒是指攻击者恶意篡改数据或向数据训练集中添加有毒数据对数据集进行污染，从而达到破坏模型和影响模型准确率的目的。因为联邦学习的训练数据只在各个客户端本地储存，对服务器和其他客户端来说都是不可见的，因此服务器无法看到客户端的训练过程。这意味着恶意客户端可以任意的毒害或篡改自己的数据而不被发现，从而生成恶意更新并上传到服务器进行聚合，从而影响整个模型的性能[7]。标签翻转攻击是数据投毒的一个典型攻击，攻击者通过改变恶意客户端的数据，使某一类的每个标签都切换成目标标签，导致模型无法正确的识别该类，影响模型的准确率[8]。除此之外，由于联邦学习的各个客户端的训练数据通常呈现非独立同分布（Non-Independent and Identically Distributed，NON-IID）的特点，NON-IID 意味着数据样本之间的分布差异较大，存在不一致和不平衡性，这导致训练过程中一些异常的局部梯度更新的存在是合理的，导致对标签翻转攻击的防御变得更加困难。

模型投毒是指攻击者破坏训练过程完整性，通过完全控制部分客户端的训练阶段,对上传的局部模型进行篡改，从而实现对全局模型的操纵。后门攻击[9]是模型投毒的主要攻击方式之一，攻击者试图使模型在某些目标任务上实现特定表现，同时保持模型在主要任务上的良好性能。由于后门仅通过特定触发器激活，因此较为隐蔽不易发现[10]。现有针对联邦学习

的后门攻击大致有两种方式，一种是集中式的后门攻击，另一种是分布式的后门攻击[11]。基于传统集中式的后门攻击没有考虑到联邦学习框架分布式的特性，攻击者使用全局后门触发器对联邦学习进行攻击，这样的后门攻击缺乏灵活性很容易被联邦学习防御机制检测到。因此向联邦学习中高效嵌入后门往往需要充分利用其分布式特征，以达到提高攻击性能、躲避防御机制的效果，但现有分布式后门攻击仅通过简单拆分全局后门触发器，这种方法具有易误触、在拜占庭鲁棒的聚合机制下攻击效果差的缺点。

因此，投毒攻击对联邦学习框架构成了严重的安全威胁，限制了联邦学习的发展和应用，研究联邦学习中投毒攻击的防御很有必要，这将提升联邦学习的健壮性，使联邦学习变得更加可靠。

2. 国内外研究现状

（1）联邦学习数据投毒攻击防御方法

联邦学习框架容易遭受恶意客户端篡改数据的投毒攻击。其中，标签翻转(label flipping)是一种典型的数据投毒攻击，它通过直接修改目标类别的训练数据的标签信息，使模型将目标标签的特征对应到错误标签，从而影响模型的准确性。由于联邦学习中服务器不能访问用户的训练数据，这使得对标签翻转攻击的防御变得更加困难，目前，有研究已经提出了许多防御标签翻转攻击威胁的策略，基于鲁棒性聚合的方法是防御标签翻转攻击比较常用的方法，大致分为基于统计分析的鲁棒性聚合方法、基于局部模型性能的鲁棒性聚合方法等。

基于统计分析的鲁棒性聚合方法将模型视为向量并利用其统计特征提取信息实现对标签翻转攻击的防御，Blanchard 等[12]提出 Krum 算法和扩展的 mutil-Krum 算法，通过计算每个模型更新与其最近更新之间欧式距离之和，选择距离之和最小的更新作为全局模型。而改进的 mutil-Krum 算法则会选择多个更新的平均值更新全局模型。Yin 等[13]提出了中值聚合和裁剪平均聚合，以每个维度为单位，选择中值或排除边缘值后的平均值作为全局模型。TOLPEGIN 等[14]利用聚类的思想，记录每个参与方的局部更新与全局更新的差值，并使用主成分分析(Principal Component Analysis, PCA)技术进行数据降维以观察正常参与方与恶意攻击者上传的更新。LI 等[15]在此基础上提出了使用 KCPA(Kernel Principal Component Analysis)和 K-means 聚类代替 PCA 的方法，从而获得更好的防御效果。但是大部分基于统计分析的鲁棒性聚合方法都需要已知恶意客户端数量的强假设，为此，文献[16]提出了通过隐马尔可夫模型估计更新质量的方法，根据中值和余弦相似性，在每次迭代中丢弃可能恶意的局部模型更新，无需恶意用户数量的假设。因此，基于统计分析的鲁棒性聚合算法计算较简单，但当统计特征、相似性的评价标准不能很好区分恶意梯度时，会极大地降低防御效果。

基于局部性能的鲁棒性算法是通过在服务器上提供的良性辅助数据集上对每个局部模型的训练优劣进行评估，依据评估结果分类聚合的权重，或者自动丢弃对准确性产生负面影响的更新。Xie 等[17]提出了使用基于得分排名机制的 Zeno 方案，该方案对每一个候选梯度都持怀疑态度，并允许任意数量的恶意用户，只需保证至少存在一个诚实用户。这类鲁棒性聚合方法直接依赖数据集的测试结果，检测结果更加的可靠，但需要预先构建好辅助数据集。

随着深度网络的兴起，对抗训练成为防御标签翻转攻击的方法之一，Shah 等[18]研究了在联邦学习环境中使用对抗训练来减少模型偏移，显著提高了对抗精度和模型收敛时间。为了防止对抗样本攻击中的逃逸攻击，Chen 等[19]通过采用高斯噪声在训练数据集中包含对抗性数据来平滑训练数据。Zhao 等[20]提出 PDGAN 方法，用生成对抗网络(Generative Adversarial Networks，GAN)生成测试数据集，用于识别数据投毒攻击，通过不断改变部署策略从而增加攻击成本和复杂度和移动目标防御(Moving Target Defense)。Shen 等[21]用 GAN 消除对抗性扰动，实现基于 GAN 的防御。但是对抗训练对于更复杂的黑盒攻击可能不具备稳定性，且加入的扰动会影响分类的精度，需要进一步采取适当的优化技术改善这些问题。

（2）联邦学习模型投毒攻击防御方法

联邦学习中的后门攻击是指恶意攻击者使模型在某些目标任务上实现特定表现，即在特定的输入下激活后门输出攻击者想要的输出，同时保持模型在主要任务上的良好性能，由于后门攻击目的通常是未知的，因此更加难以被检测。

目前，在联邦学习中，结合模型投毒的后门攻击更为常见，因此也可把后门攻击称为有针对的模型中毒，主要分为标记后门攻击[22]和语义后门攻击[23]两种，Wu 等[24]证明了基于特定标记的后门攻击在数据非独立同分布程度越高时攻击越有效。无论哪种攻击方式，后门攻击的效果只在特定的输入才会触发后门。Zhou 等[25]提出了基于优化模型的后门攻击， 通过将冗余神经元训练为对抗神经元来实现攻击。该文献通过实验证明了这种后门攻击不仅能实现较高的攻击成功率，还能规避一些防御措施。Wang 等[26]研究了对抗样本攻击与后门攻击之间的联系，表明模型对后门的鲁棒性在通常情况下意味着对于对抗样本攻击的鲁棒性。Xie 等[27]提出了新的分布式后门攻击，即后门在恶意攻击者控制的用户之间被拆分，并将每个模式嵌入敌对客户的训练集中，在模型聚合后又将成为一个完整的后门并插入到模型中，从而提高后门攻击的隐蔽性。并且许多研究已经证明了后门攻击凭借其隐蔽性，能够在仅发动攻击成功一次的情况下，使得全局模型能在多轮的迭代中保留后门。另外，鉴于联邦学习通常采用安全聚合算法[28]，中心服务器无法检查参与者的参数更新，即无法检测参与者对全局模型的异常贡献，这使得防御后门攻击成为一个难点。针对后门攻击，目前已提出了多种不同的防御方法。Sun 等[29]采用了梯度剪裁与添加噪声的方法，通过将参与者更新范数限制在阈值范围

之内，并对剪裁后的全局模型添加高斯噪声，从而减轻了恶意参与者对全局模型的影响。然而，若梯度剪裁的值过于宽松，则无法防御后门攻击，但过于严苛会导致模型无法收敛。Gao等[30]采用了限制参与方更新上传比例的聚合规则，该规则要求参与方随机选择部分参数上传至中心服务器，同时设计了新的安全聚合协议，以便服务器检查参与方是否上传了部分参数而非全部参数。然而，Li 等[31]验证了攻击者可以通过上传中毒的神经通路来完成对联邦模型的后门攻击，因此由参与方选择上传参数的方法并不能完全消除后门攻击。

综上针对联邦学习的投毒攻击防御方法存在以下问题：

（1）联邦学习框架面临的数据投毒中的标签翻转攻击时，由于联邦学习客户端的数据具有不平衡性，因此用户上传的局部模型参数具有多样性，使用单纯的鲁棒性防御方法判断局部更新是否为恶意更新变得更加困难，从而难以保证防御的效果。

（2）现有联邦学习所面临的基于后门触发器并修改模型参数的后门攻击威胁由于攻击方式更加灵活隐蔽，变得更加难以检测的问题。

（3）现有联邦学习基于触发器的数据投毒后门攻击的防御方法中在客户端无法获得其他客户端的更新情况下，如何从中识别出离群的恶意模型更新并清除后门的问题。

二、论文的研究内容、研究目标，以及拟解决的关键问题（包括具体研究与开发的主要内容、目标和要重点解决的关键技术问题）

论文主要对联邦学习框架所面临的投毒攻击与防御方式进行深入的研究，拟从标签翻转攻击的防御方法、基于随机断层与梯度裁剪的联邦学习后门攻击防御、基于 Grad-CAM 的联邦学习后门攻击防御方法设计这三个方面作为研究内容。针对联邦学习由于数据分布不平衡，局部梯度差异大，容易遭受中毒攻击，且目前防御机制不易识别更新是否为恶意更新，因此针对标签翻转攻击设计基于辅助训练集的防御方法。针对具有后门触发器修改模型参数的模型中毒后门攻击检测困难的问题，提出一种基于随机断层与梯度剪裁相结合的后门防御策略和技术方案。针对现有联邦学习框架加入同态加密等安全聚合机制后，传统的防御方法效果不佳的问题，提出基于 Grad-CAM 的联邦学习后门攻击防御方法。

1.主要研究内容及研究目标

（1）基于辅助训练集的标签翻转攻击防御

针对联邦学习的标签翻转攻击问题，提出了基于辅助训练集的标签翻转攻击防御方法。首先构建攻击者模型，通过发起标签翻转攻击，降低模型的准确率，然后通过生成对抗网络重建辅助数据集，利用暴露的特征分布和其余大多数自然对抗性扰动之间的基本统计异质性，

对辅助训练集进行标签分类，最后基于辅助训练集，对全局模型再训练，中和标签翻转攻击的影响，提升模型准确率。

（2）基于随机断层与梯度裁剪的联邦学习后门攻击防御

针对联邦学习所面临的后门威胁，从博弈的角度，提出一种基于随机断层与梯度剪裁相结合的后门防御策略和技术方案。中心服务器在收到参与方提交的梯度信息后，随机确定每个参与方的神经网络层，然后将各参与方的梯度贡献分层聚合，并使用梯度阈值对梯度参数进行裁剪。梯度剪裁和随机断层可削弱个别参与方异常数据的影响力，使联邦模型在学习后门特征时陷入平缓期，长时间无法学习到后门特征，同时不影响正常任务的学习。如果中心服务器在平缓期内结束联邦学习，即可实现对后门攻击的防御。

（3）基于 Grad-CAM 的联邦学习后门攻击防御

针对现有后门攻击的防御方法在联邦学习添加安全聚合机制后效果变得不明显的问题，提出了一个运行在客户端本地的，不改变现有的联邦学习架构的防御方法。客户端将另外建立一个与全局模型结构一致的模型，称为评估模型，在使用私有数据训练本地模型的同时，也会训练评估模型。评估模型不会像本地模型一样接收来自全局模型的更新，是完全用私有数据训练出的模型。当用户数据中带有后门触发器时，程序可以将后门删除，确保全局模型总是收到干净的输入。并且利用 GAN 模型学习无触发器图像的分布，以进行干净数据的恢复。

2.拟解决的关键问题

1)针对联邦学习数据中毒中的标签翻转攻击现有鲁棒性聚合防御方法不能很好判断局部梯度是否是恶意更新，对有些看似是恶意更新实则是正常更新的局部模型参数直接过滤，导致防御效果下降的问题。设计了一种基于辅助训练集的防御方法提高防御效果。

2)针对联邦学习基于模型中毒后门威胁的现有防御策略在抵御后门攻击时表现不佳问题的问题，从博弈的角度，提出一种基于随机断层与梯度剪裁相结合的后门防御策略和技术方案。

3)针对现有联邦学习框架在加入同态加密等安全聚合机制后，对数据投毒的后门攻击防御方法中如何从中识别出离群的恶意模型更新并清除后门的问题，提出基于 Grad-CAM 的联邦学习后门攻击防御方法

三、拟采取的研究方案及可行性分析(包括研究的基本思路，研究过程拟采用的方法和手段，现有研究条件和基础，研究开发方案和技术路线等)

针对以上的研究内容及拟解决的关键问题，首先研究基于辅助训练集的标签翻转攻击防御方法、基于随机断层与梯度裁剪的联邦学习后门攻击防御方法和基于 Grad-CAM 的联邦学习

后门攻击防御方法。

1、 基于辅助训练集的标签翻转攻击防御

相比于后门攻击需要设计特定的触发器生成算法和训练后门样本，标签翻转攻击作为一种轻量级攻击方法，只需要翻转部分数据集标签即可大幅度降低模型的分类精度，带来严重后果。由于联邦学习各个用户的数据集基本基本上是非 NON-IID 的，因此目前的鲁棒性聚合防御机制依靠过滤与全局模型相违背的梯度效果不佳，而利用辅助训练集能够直接依赖数据集的测试结果，防御效果更好，但构建辅助数据集较为困难。

针对联邦学习数据中毒中的标签翻转攻击的防御问题，设计基于辅助训练集的方法抵御标签翻转攻击。具体流程如下图 3-1 所示

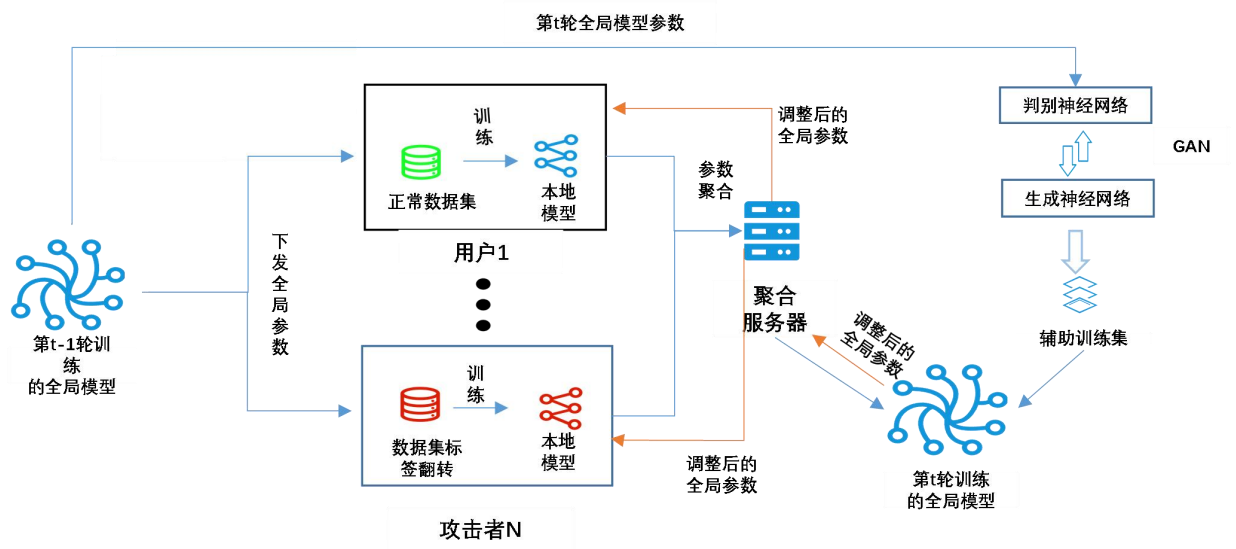

图 3-1　基于辅助训练集的标签翻转攻击框架图

具体流程：

（1）投毒阶段。随机选择若干个客户端作为恶意客户端，对恶意客户端的本地训练数据集的某类数据标签翻转成目标标签，将本地训练数据修改成毒化数据进行模型训练，上传毒化后的本地模型参数，进而毒化整个全局模型。

（2）构建辅助训练集阶段。利用 GAN 生成辅助数据集，并对辅助数据集进行标签分类。将联邦学习上一轮的全局模型参数和 GAN 生成神经网络生成的伪样本作为 GAN 判别神经网络的输入，在经过 GAN 生成神经网络和判别神经网络的单独交替迭代训练后，最终从 GAN 生成器中得到符合条件的输出，即辅助数据集。

（3）全局模型再训练阶段。使用 GAN 构建并且分类好标签的辅助训练集对上一轮全局模型进行训练，生成新的全局模型参数，新的全局模型参数代替原本的用户平均模型参数，并分发给特定的良性用户，继续进行迭代训练学习，从而达到抵御投毒攻击的效果。

2、 基于随机断层与梯度裁剪的联邦学习后门攻击防御

针对模型投毒的后门攻击，提出一种随机断层与梯度裁剪相结合的更新策略，在神经网络进行梯度更新时，随机断层要求神经网络随机抛弃一部分神经网络层的参数，只对保留参数的神经网络层进行更新，并对其进行梯度裁剪，该策略会使神经网络训练过程产生数轮的平缓期，使其特征学习变慢、中心服务器将每个参与者提交的参数与随机生成的掩码相乘，实现随机保留部分神经网络层的参数，并在梯度分层聚合结束后对梯度更新进行剪裁。由于联邦学习中正常参与者始终占多数，正常参与者之间可以实现梯度贡献的互补，联邦模型训练目标任务过程不会产生平缓期，也不会影响模型的整体收敛性；而联邦学习中攻击者占少数，后门攻击者训练后门任务时更像是在训练单个神经网络模型，随机断层与梯度剪裁相结合的防御策略会使后门训练进入数轮的平缓期，从而抑制了联邦学习过程中可能存在的后门攻击，也保证了目标任务的训练。具体步骤流程如图 3-2 所示

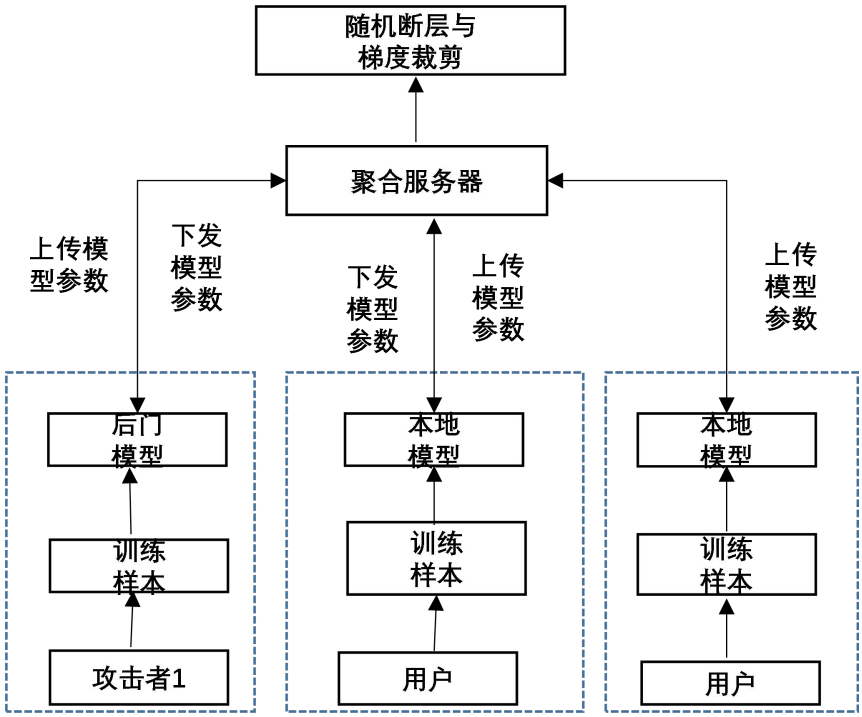

图 3-2 基于随机断层与梯度裁剪的联邦学习后门攻击防御框架图

具体流程如下：

（1）初始化。中心服务器初始化联邦模型参数并分发给所有参与者。

（2）训练防御。中心服务器按照采样率随机选取每轮联邦学习的参与者集合，参与方执行本地模型训练后将梯度值进行同态加密后上传至中心服务器，中心服务器为每个参与方随机选取神经网络层更新 0-1 掩码矩阵，并在同态加密下计算掩码矩阵和，然后中心服务器根据该层参与更新的用户数计算分层梯度平均聚合，最后在中心服务器执行设定的阈值进行个体梯度裁剪。

（3）分发梯度。中心服务器向所有的参与方分发最后的梯度更新值。

3、 基于 Grad-CAM 的联邦学习后门攻击防御

针对防御模块的核心是在客户端本地实现对训练数据进行检测和处理，从而达到防止后门攻击的目的。其中共分为三个阶段，具体流程图如下图 3-3 所示

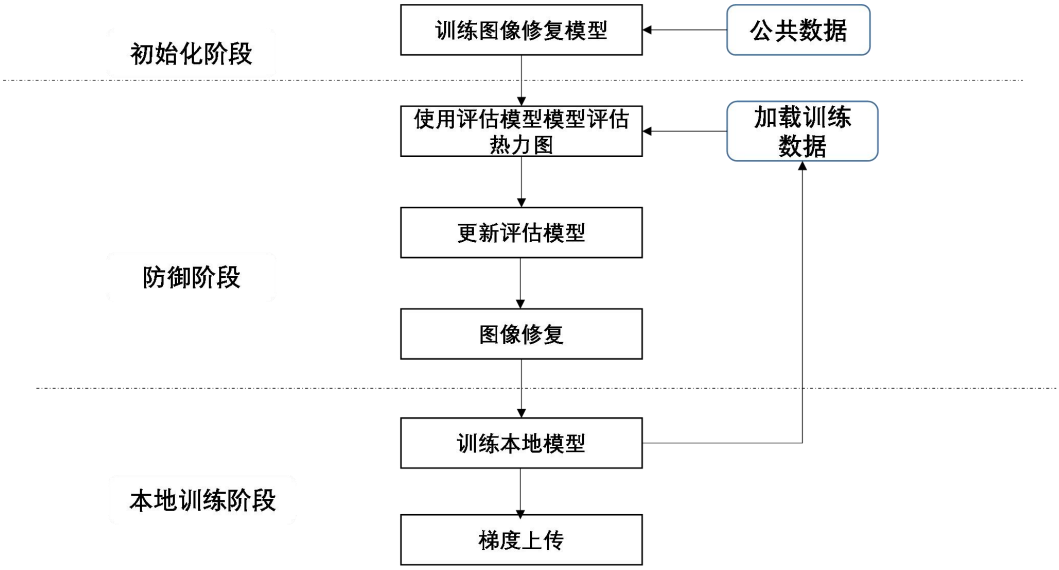

图 3-3 基于 Grad-CAM 的联邦学习后门攻击防御方法框架图

具体步骤如下：

（1）初始化阶段。该阶段主要进行的是防御框架的准备和部署。由于在后续的图像修复过程中需要用到图像修复模型，所以在联邦学习还未开始前，要先对该模型进行训练。模型训练可以由服务器进行，再分发部署到本地客户端上。服务器可以使用全局任务的部分测试集，或者与测试集具有相同分布的无标签数据作为图像修复模型的训练数据。

（2）防御阶段。该阶段主要进行的是后门数据的识别与修复以及评估模型的更新。为了使 Grad-CAM 可以识别出后门区域，首先需要一个嵌入了后门的模型，称为评估模型。首先使用用户的训练数据单独训练一个模型，将其作为评估模型。在每轮迭代中，使用同一批次数据对评估模型与联邦学习本地模型进行训练。如果用户的训练数据集含有后门样本，由于触发器模式的简单性和直接性，评估模型会很快学习到后门模式。防御阶段的流程为：先计算热力图，输出遮盖图像，之后更新评估模型，最后将遮盖图像输入修复模型进行修复。

（3）本地训练阶段。将修复后的图像输入到本地模型进行训练，本地模型训练完成后将梯度上传到参数服务器，这样可以保证全局模型不受基于数据投毒的后门攻击的影响。

四、本课题的特色与创新之处

（1）针对数据投毒中的翻转标签攻击防御中的鲁棒聚合方法中构建辅助数据集进行防御比较困难，将生成对抗方法引进鲁棒聚合方法中，利用 GAN 模型生成辅助数据集进行清洗有

毒数据集。

（2）提出一种应对植入触发器后门攻击的神经网络随机断层更新策略，能够在联邦学习框架添加同态加密机制下仍然具有很好的防御效果。

（3）针对数据投毒的后门攻击，在不改变联邦学习现有框架下，将防御手段由服务器转向客户端，在客户端使用 Grad_CAM 检测并定位后门触发器，然后使用基于 GAN 的图像修复方法对后门触发器进行清除，恢复干净样本。

五、参考文献

<页面、页数不足请自行加页>

[1] Li J. Cyber security meets artificial intelligence: A survey. Frontiers of Information Technology & Electronic Engineering, 2018, 19(12):1462-1474.

[2] 周传鑫, 孙奕, 汪德刚, 葛桦玮. 联邦学习研究综述[J]. 网络与信息安全学报, 2021, 7(5):77-92.

[3] Konečný J, McMahan H B, Ramage D, et al. Federated optimization: distributed machine learning for on-device intelligence. arXiv preprint arXiv:1610. 02527, 2016.

[4] McMahan B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data//Artificial intelligence and statistics. Florida, USA, 2017: 1273-1282.

[5] SUN G, CONG Y, DONG J, et al. Data poisoning attacks on federated machine learning[J]. IEEE Internet of Things Journal, 2021, 9(13):11365-11375.

[6] CAO X, GONG N Z. Mpaf: Model poisoning attacks to federated learning based on fake clients[C]// Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022:3396-3404.

[7] 高莹,陈晓峰,张一余等.联邦学习系统攻击与防御技术研究综述[J].计算机学报,2023,46(09).

[8] 陈学斌,任志强,张宏扬.联邦学习中的安全威胁与防御措施综述[J].计算机应用.

[9] 王永康,翟弟华,夏元清.联邦学习中抵抗大量后门客户端的鲁棒聚合算法[J].计算机学报,2023,46(06).

[10] 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).

[11] 林智健.针对联邦学习的组合语义后门攻击[J].智能计算机与应用,2022,12(07).

[12] Blanchard P, El Mhamdi E M, Guerraoui R, et al. Machine learning with adversaries: Byzantine tolerant gradient descent//Proceedings of the International Conference on Neural Information Processing Systems. Long Beach, USA, 2017: 118－128.

[13] Yin D, Chen Y, Kannan R, et al. Byzantine-robust distributed learning: Towards optimal statistical rates//Proceedings of the International Conference on Machine Learning. Vienna, Austria, 2018: 5650-5659.

[14] TOLPEGIN V, TRUEX S, GURSOY M E, et al. Data poisoning attacks against federated learning

systems[C]// Proceedings of the 2020 European Symposium on Research in Computer Security. Cham: Springer, 2020: 480-501.

[15] LI D, WONG W E, WANG W, et al. Detection and mitigation of label-flipping attacks in federated learning systems with KPCA and K-means[C]// Proceedings of the 2021 International Conference on Dependable Systems and Their Applications (DSA). Piscataway: IEEE, 2021: 551-559.

[16] Muñoz-González L, Co K T, Lupu E C. Byzantine-robust federated machine learning through adaptive model averaging. arXiv preprint arXiv:1909.05125, 2019.

[17] Xie C, Koyejo S, Gupta I. Zeno: Distributed stochastic gradient descent with suspicion-based fault-tolerance//Proceedings of the International Conference on Machine Learning. California, USA, 2019: 6893-6901.

[18] Shah D, Dube P, Chakraborty S, et al. Adversarial training in communication constrained federated learning. arXiv preprint arXiv:2103. 01319, 2021.

[19] Chen C, Kailkhura B, Goldhahn R, et al. Certifiably-Robust Federated Adversarial Learning via Randomized Smoothing//Proceedings of the IEEE International Conference on Mobile Ad Hoc and Smart Systems. Denver, USA, 2021: 173-179.

[20] Zhao Y, Chen J, Zhang J, et al. PDGAN: A novel poisoning defense method in federated learning using generative adversarial network //Proceedings of the International Conference on Algorithms and Architectures for Parallel Processing. Melbourne, Australia, 2019: 595-609.

[21] Shen S, Jin G, Gao K, et al. Ape-gan: Adversarial perturbation elimination with gan//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton, UK, 2019: 3842-3846.

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

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

[24] WU C, YANG X, ZHU S, et al. Mitigating backdoor attacks in federated learning[EB/OL]. (2021-01-14) [2023-07-09].

[25] ZHOU X, XU M, WU Y, et al. Deep model poisoning attack on federated learning[J]. Future Internet, 2021, 13(3): 73.

[26] Wang H, Sreenivasan K, Rajput S, et al. Attack of the tails: Yes, you really can backdoor federated learning//Proceedings of the International Conference on Neural Information Processing Systems. Virtual, 2020: 16070-16084.

[27] Xie C, Huang K, Chen P Y, et al. Dba: Distributed backdoor attacks against federated

learning//Proceedings of the International Conference on Learning Representations. New Orleans, USA, 2019.

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

[29] SUN Z, KAIROUZ P, SURESH A T, et al. Can you really backdoor federated learning?[EB/OL]. (2019-12-02) [2023-07-09].

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

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

[32]张思思,左信,刘建伟.深度学习中的对抗样本问题[J].计算机学报,2019,42(08):1886-1904.