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。

2. 国内外研究现状

（1）基于特定样本触发器的联邦学习隐式后门攻击

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。

（1）基于 DCGAN 和特征的联邦学习后门攻击

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，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。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。针对后门攻击，目前已提出了多种不同的方法，Bagdasaryan 等[24]提出了一种基于模型替换的投毒方法，其依据模型收敛性导致局部模型更新趋于零，利用模型替换在一轮迭代中将全局模型替换为后门模型。Wu 等[25]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。文献[23]提出了规避防御的语义后门攻击，并通过放大本地更新实现模型替换，同时也证明了攻击效果会随着诚实参与者的参与而下降。但是，这一方法需要攻击者熟知异常检测策略这一强假设。为此 Sun 等[26]提出了范数有界攻击(Norm Bounded Backdoor Attack)，通过将更新进行约束以规避一些防御措施，并证明了其攻击的有效性。同时还量化了恶意攻击者的人数与参与攻击的频率对后门攻击的影响，证明了攻击者比例越大，攻击者参加频率越高则后门攻击的效果越好。

由于对抗样本攻击在机器学习中的应用比较广泛，随着联邦学习的发展，在联邦学习场景下的模型部署与机器学习类似，传统的对抗样本攻击可以拓展到联邦学习中，Zhou 等[27]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

[28]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。Pang 等[29]则针对纵向联邦学习中用户的特征差异，提出了对抗性主导输入攻击（Adversarial Dominating Inputs Attack)。与传统的对抗性样本控制整个特征空间不同，对抗性主导输入攻击仅仅控制部分特征输入，就能主导其他用户的全部输入，实现对特定的输入进行错误分类。 同时，对抗性主导输入使得其他用户做出非常少的贡献，从而影响了激励用户贡献的奖励。

（3）基于组合语义特征的联邦学习分布式后门攻击

目前，大多数后门攻击以集中式后门为主，即攻击者使用全局触发器对联邦学习进行攻击，并没有考虑联邦学习分布式特点[30]，也很容易被联邦学习防御机制过滤。Baruch 等[31]发现如果允许攻击者串通共谋，投毒攻击的效果会大大提高，这种勾结可以让对手创建模型更新攻击，既更有效，也更难以发现。因此，Xie 等[32]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。并证明了相较于集中式的后门攻击，联邦学习遭受到分布式后门攻击的危害更大。但是这种方法可能需要攻击者之间预先协商全局后门触发器，并且这种攻击方法的误触率很高，局部触发器很容易触发后门分类，也很容易在拜占庭聚合方法下被过滤，因此研究出更灵活和隐蔽的分布式后门攻击方式值得深入研究。

综上针对联邦学习的投毒攻击与防御方法存在以下问题：

（1)针对联邦学习投毒攻击中的后门攻击以水印方式植入的后门触发器与原样本差异较大，易被鲁棒性聚合机制检测发现并过滤导致攻击失败的问题。

（2)针对现有联邦学习投毒攻击中的后门攻击中给攻击者的训练样本中植入后门触发器时，每张图片植入的后门触发器标记相同，可能会被检测到甚至是重建出后门触发器从而导致攻击失败的问题。

（3）现有联邦学习分布式后门攻击只是简单的将后门触发器进行拆分，具有较高的误触率，同时也容易被鲁棒性聚合方法检测到并过滤，因此设计出隐蔽和灵活的且不易被鲁棒性聚合方法过滤的分布式后门攻击方式是值得考虑的问题。

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，拟从基于 DCGAN 和特征的联邦学习后门攻击、基于特定样本触发器的联邦学习隐式后门攻击、分布式后门攻

击方法设计这三个方面作为研究内容。

1.主要研究内容及研究目标

（1）基于 DCGAN 和特征的联邦学习后门攻击

针对联邦学习投毒攻击中的后门攻击中的触发器与干净样本梯度差异较大而易被联邦学习防御机制检测并过滤，以及后门攻击收敛速度较慢的问题，提出了 DCGAN 和特征的后门攻击方案，首先利用 DCGAN 模型生成伪样本，然后反向输入到全局模型中进行预测，选择准确率最高的样本作为触发器，通过特征方式植入后门触发器，降低触发器样本与原样本的梯度差异，然后缩放后门模型加快模型收敛速率，从而提升后门攻击成功率。

（2）基于特定样本触发器的联邦学习隐式后门攻击

针对联邦学习投毒攻击中的后门攻击中的每个样本植入标记相同的触发器容易被检测甚至重建的问题，提出了基于特定样本触发器的联邦学习隐式后门攻击。首先构建攻击者模型，通过图像隐写网络给给攻击者的训练样本写入相同的隐式信息，但是触发器不一样，然后进行训练本地后门模型，通过模型缩放保持后门的攻击效果。

（3）基于组合语义特征的联邦学习分布式后门攻击

联邦学习分布式后门攻击的攻击高误触率和易被联邦学习拜占庭聚合机制过滤，设计一种针对组合语义特征的联邦学习分布式后门攻击方法，首先攻击者们独立选择一个已有标签的对象的语义特征作为局部语义触发器，然后选择临时特征插入到本地训练样本中，训练本地局部后门模型上传至服务器，组合构成全局组合语义触发器。当来自多个标签的组合语义特征全部出现触发后门输出目标标签。攻击过程中无需知识共享，减轻了局部触发器误触的发生，这种后门攻击方式更加灵活自然且不易被检出。

2.拟解决的关键问题

1）针对联邦学习投毒攻击中的后门攻击以水印方式植入的后门触发器与原样本差异较大，易被鲁棒性聚合机制检测发现并过滤导致攻击失败的问题。

2）针对现有联邦学习投毒攻击中的后门攻击中给攻击者的训练样本中植入后门触发器时，每张图片植入的后门触发器标记相同，可能会被检测到甚至是重建出后门触发器从而导致攻击失败的问题。

3）由于联邦学习集中式后门攻击使用全局后门触发器易被检测器检测，而基于像素的分布式后门攻击嵌入不属于任何输出标签的新特征容易在训练过程中是模型的参数修改幅度较大，容易被防御机制判定为异常更新从而过滤，使后门攻击不成功。因此设计一种更加灵活隐蔽的基于组合语义特征的分布式后门攻击方法。

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 GAN 的联邦学习集中式后门攻击方法和基于组合语义特征的联邦学习分布式后门攻击方法。

1、基于 DCGAN 和特征的联邦学习后门攻击

由于联邦学习框架中，攻击者无法得到其他用户的训练数据，因此传统的后门触发器往往是由像素点或者图片组成，导致后门触发器样本与原样本梯度差异较大容易被防御机制检测到并过滤。并且联邦学习在聚合的过程中使用参数平均放方法，这可能会使后门攻击的贡献在聚合的过程中被抵消，导致后门攻击需要多轮训练才能攻击成功，后门模型的收敛速度较慢。因此论文拟采用 DCGAN 和特征对传统的后门触发器进行改进，同时使用模型缩放技术加快模型模型收敛速度，使后门攻击更加灵活隐蔽，具体步骤流程如图 3-2 所示

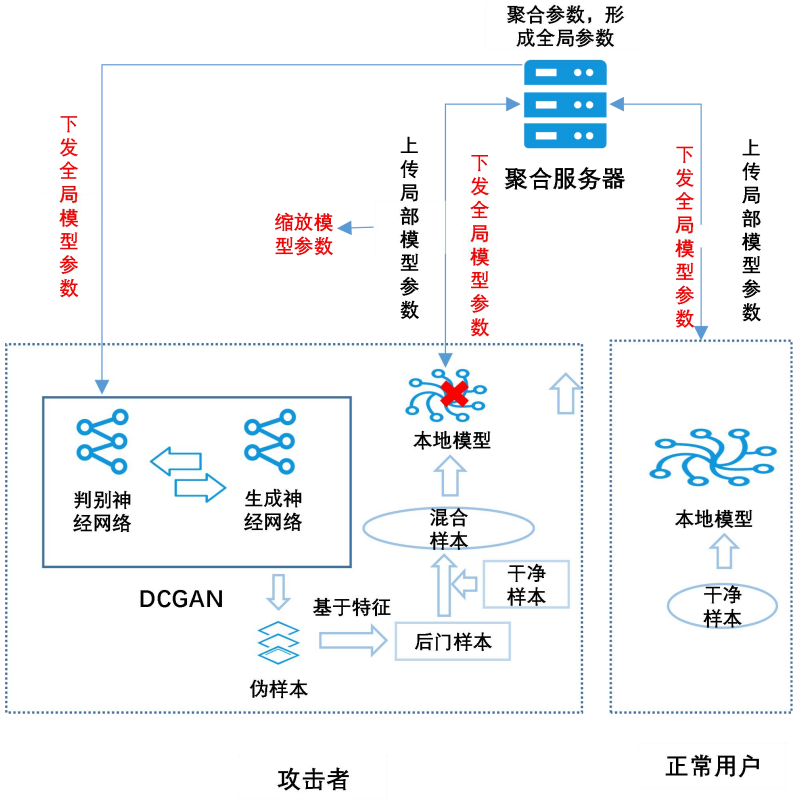

图 3-2 基于 DCGAN 和特征的联邦学习后门攻击框架图

具体流程如下：

（1）正常训练。攻击者首先伪装成良性客户端，下载全局模型参数，更新本地模型和判别器模型，并上传模型参数。

（2）后门触发器生成。在正常训练的过程中，利用 DCGAN 获取其他用户的真实训练样本，通过生成器生成伪样本，然后将伪样本和模型全局参数输入判别器，判断是否属于标签空间，经过不断迭代，得到其他用户的训练样本。然后将训练样本反输入到全局模型当中进行预测，选择准确率最高的样本类别作为后门触发器。

（3）后门样本生成和训练。将生成的触发器样本以特征的方式引入入到攻击者的本地训练样本中形成后门样本，利用混合良性特征训练即将后门样本，干净样本、触发器样本形成混合样本共同训练本地模型，从而引入后门。

（4）模型缩放。攻击者训练后门模型后，将模型参数进行缩放并上传至服务器，使后门模型在聚合的过程中不会抵消其贡献，最终替换全局模型。

（5）扩展分析存在多个攻击者的情形。

2、 基于特定样本触发器的联邦学习隐式后门攻击

针对针对现有联邦学习投毒攻击中的后门攻击中给攻击者的训练样本中植入后门触发器时，每张图片植入的后门触发器标记相同，可能会被检测到甚至是重建出后门触发器从而导致攻击失败的问题，设计基于特定样本触发器的联邦学习隐式后门攻击。具体流程如下图 3-1 所示

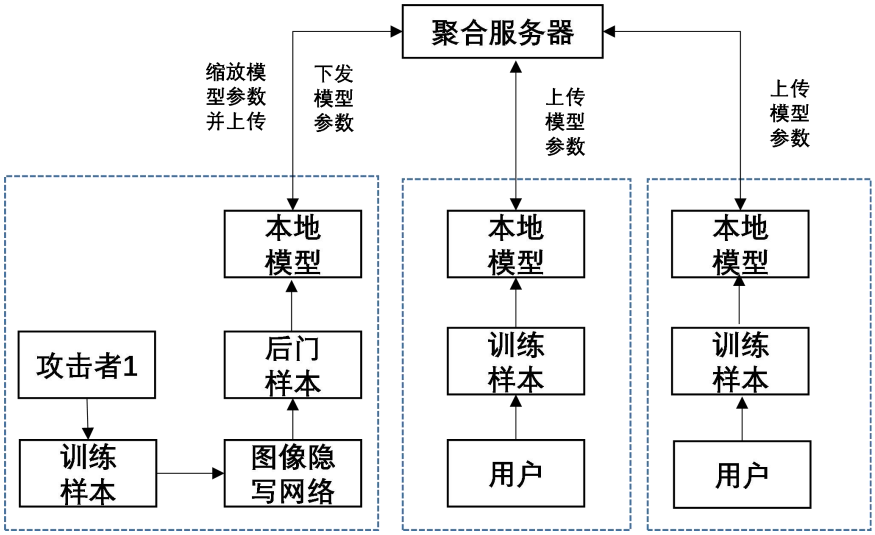

图 3-1　基于特定样本触发器的联邦学习隐式后门攻击框架图

具体流程：

（1）攻击者伪装成正常用户参与训练直至模型接近收敛。

（2）通过图像隐写网络对攻击者的样本隐写进同样信息，并修改标签得到后门样本。

（3）由于图像隐写网络的特殊性，每一个样本被植入的触发器都不同，实现了样本特定触发器的设定。

（4）攻击者利用后门样本训练后门模型，并对模型参数进行缩放上传聚合服务器。

（5）扩展分析存在多个攻击者的情形。

3、 基于组合语义特征的联邦学习分布式后门攻击

分布式学习在机器学习中被广泛应用，现有的联邦学习集中式后门攻击没有充分利用联邦学习框架分布式特点，攻击者使用全局后门触发器容易被防御机制检测到，且目前的联邦学习的分布式后门攻击多是基于像素特征的触发器，将一个基于像素的全局触发器简单分散给各个攻击者，每个攻击者拥有一个局部的触发器，经过训练后将各自的局部触发器上传至服务器组成一个全局后门触发器。由于基于像素的触发器需要嵌入到数据样本中，嵌入的像素特征会成为目标标签的一个强大特征，对模型参数的修改较大，容易被鲁棒性聚合机制检测并过滤。本文拟设计一种基于语义特征的联邦学习分布式后门攻击方法，利用语义后门不用引入其他独特特征的特点，将来自多个标签的语义特征进行组合形成后门触发器，使攻击变得更加灵活和隐蔽，具体流程图如下图 3-3 所示

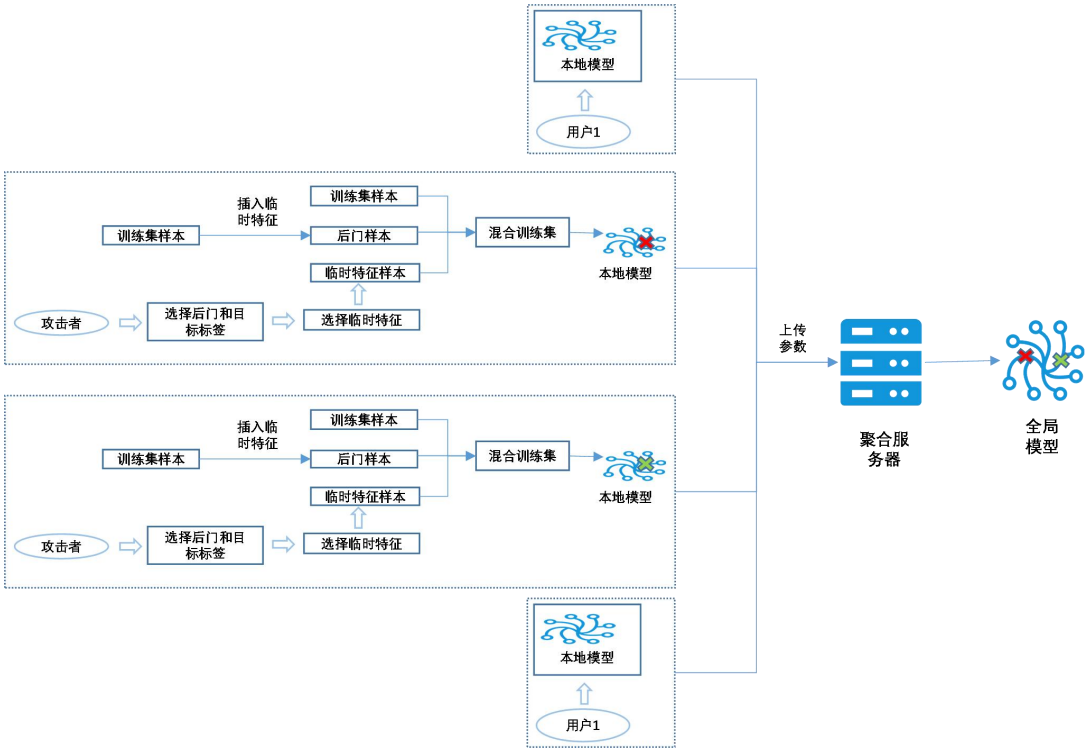

图 3-3 基于组合语义特征分布式后门攻击框架图

具体步骤如下：

（1）攻击者制定后门特征和目标输出。攻击者们分别独立选择一个已有标签的对象的语义特征作为局部触发器，并且希望当所有被选择的良性对象的语义特征同时出现时触发全局后门，输出攻击者的目标标签。

（2）训练生成局部触发器。攻击者确定局部触发器对象后，将临时特征插入到局部触发

器对象对应的训练样本数据中，并修改他们的标签为目标标签，最终新训练集包括原始正常样本、后门样本以及消除临时特征点样本。能够让模型在良性样本输入时分类正常，而当攻击者使用带触发器的样本时，样本会被模型分类为特定的错误类别。

（3）联邦学习聚合过程。利用联邦学习的聚合过程生成全局后门模型，局部后门模型在参数服务器的聚合阶段进行组合，最终生成全局后门模型，当全局触发器出现时触发恶意后门分类。

四、本课题的特色与创新之处

（1）利用图像隐写网络植入相同信息不同形状的触发器，防止良性用户根据后门模型推导出后门触发器，使后门攻击更加灵活隐蔽，扩展分析有多个攻击者进行攻击的情况。

（2）利用图像隐写网络植入相同信息不同形状的触发器，防止良性用户根据后门模型推导出后门触发器，使后门攻击更加灵活隐蔽，扩展分析有多个攻击者进行攻击的情况。

（3）相比集中式后门攻击的全局后门触发器和注入不属于任何输出标签的新特征的基于像素的分布式后门攻击，使用已有标签的多个对象的组合语义特征作为后门触发器，发起分布式后门攻击，更加灵活自然，攻击方式也更加隐蔽。

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] Bagdasaryan E, Veit A, Hua Y, et al. How to backdoor federated learning //Proceedings of the International Conference on Artificial Intelligence and Statistics. Virtual, 2020: 2938-2948.

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

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

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

[28] 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.

[29] Pang Q, Yuan Y, Wang S. Attacking Vertical Collaborative Learning System Using Adversarial Dominating Inputs. arXiv preprint arXiv:2201.02775, 2022.

[30] Chen Z, Tian P, Liao W, et al. Towards multi-party targeted model poisoning attacks against federated learning systems. High-Confidence Computing, 2021, 1(1): 100002.

[31] Baruch M , Baruch G , Goldberg Y. A Little Is Enough: Circumventing Defenses For Distributed Learning[J]. 2019.

[32] 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.