基于联邦学习的分布式后门攻击方法研究

向舒悦 [1)]

[1)](南京邮电大学, 计算机学院、软件学院、网络空间安全学院, 南京, 中国, 210046)

摘　要　联邦学习（Federated Learning）作为一种分布式学习框架，允许多个参与方在不共享原始数据的情况下共同训练模型，从而实现数据隐私的保护和模型质量的提升。然而，其分布的特性也隐藏着巨大的威胁，后门攻击就是其中之一。现有的研究大多关注于单一客户端实施后门攻击，并未结合联邦学习的分布式特性，因此本文分析了现有联邦学习后门攻击方法的局限性，提出了分布式后门攻击（Distributed Backdoor Attack，DBA），充分利用联邦学习的分布式特性，将全局触发器分解为独立的局部触发器，并将它们分别嵌入到不同攻击者的训练集中。经过大量的实验验证，本文发现这种分布式后门攻击方法不仅具有更高的持久性和隐蔽性，其攻击成功率也明显高于传统的集中式后门攻击。为了进一步探索 DBA 的特性我们还通过改变不同的触发因素来测试攻击性能，且与单触发模式的攻击效果进行了对比，表明了本文方法的优越性。此外，本文还探讨了常见的联邦学习中后门攻击的防御策略对分布式后门攻击的防御性能，包括鲁棒聚合等方法。通过分析 DBA 的优势和防御策略的优缺点，本文对联邦学习安全性的提供了更深刻的理解，并为未来的研究方向提供了有益的参考。

关键词　联邦学习；后门攻击；集中式后门攻击；分布式后门攻击

Research on Distributed Backdoor Attack Methods Based on Federated Learning

Shuyue Xiang[1)]

[1)](School of Computer Science and Technology, School of Software, School of Cyberspace Security, Nanjing University of Posts and Telecommunications, Nanjing, China，210046

Abstract

Federated Learning, as a distributed learning framework, allows multiple participants to jointly train models without sharing raw data, thereby protecting data privacy and improving model quality. However, the nature of its distribution also hides a huge threat, and backdoor attacks are one of them. Therefore, this paper analyzes the limitations of the existing federated learning backdoor attack methods, and proposes a distributed backdoor attack (DBA), which makes full use of the distributed characteristics of federated learning to decompose global triggers into independent local triggers and embed them into the training sets of different attackers. After a large number of experiments, this paper finds that this distributed backdoor attack method not only has higher persistence and concealment, but also has a significantly higher attack success rate than the traditional centralized backdoor attack. In order to further explore the characteristics of DBA, we also test the attack performance by changing different triggers, and compare the attack effect with the single-trigger mode, which shows the superiority of the proposed method. In addition, this paper also discusses the defensive performance of common backdoor attacks in federated learning against distributed backdoor attacks, including robust aggregation and other methods. By analyzing the advantages of DBA and the advantages and disadvantages of defense strategies, this paper provides a deeper understanding of federated learning security and provides a useful reference for future research directions.

Keywords　federated learning; backdoor attacks; centralized backdoor attacks; distributed backdoor attacks

1 绪论

1.1 研究背景及意义

随着大数据时代的到来，数据的获取、存储和处理已成为现代科技发展的重要基石。然而，数据的安全性问题也日益凸显，成为制约数据共享和高效利用的关键因素。在这一背景下，联邦学习 [1] 作为一种新兴的分布式机器学习技术，允许多个参与方在无需直接共享原始数据的前提下，共同协作训练一个模型。在这种范式下，数据无需集中到一个中心节点进行统一训练，而是利用各个参与方的本地设备独立进行模型训练，并将训练结果安全地汇总至中心节点，以优化整体模型性能 [2]。这一过程实现了数据隐私保护和数据共享计算的平衡，为解决数据隐私和安全性问题提供了新的思路。

尽管联邦学习在保护数据隐私方面具有显著优势，但其也面临着诸多挑战，后门攻击就是其中之一，它可能导致模型的不安全性和模型失效。后门攻击是在深度神经网络（Deep Neural Networks，DNN）模型的训练阶段，攻击者通过注入恶意样本，以隐匿的方式改变模型训练后的性能而不被察觉 [3]，使后门模型在良性样本上表现良好，而在后门数据上产生错误输出。这种策略旨在绕过 DNN 安全系统的防御机制，是一种专门针对网络安全的新型攻击技术。在联邦学习的场景中，后门攻击表现为恶意攻击者独立地在本地环境中训练出一个带有后门的模型，随后通过特定设计的模型更新策略，将后门触发器（Backdoor Trigger）嵌入到全局模型中，确保在特定触发条件的作用下，该模型能够执行预期的恶意攻击行为 [4]。这种攻击不仅破坏了模型的准确性和可靠性，还可能泄露用户隐私，对联邦学习系统的安全性造成严重威胁。

联邦学习场景中的后门攻击方法不仅隐蔽性强，而且难以被检测和防御，为了应对这种安全性挑战，研究者们提出了包括以“数据清洗”、“模型清洗”、“模型剪枝”为原理的防御方法 [5]，一方面，通过不断探索新的后门攻击方法，发现联邦学习系统中的潜在漏洞，以提出新的防御方法，另一方面，通过积极研究有效的防御策略，以提高联邦学习系统的安全性和鲁棒性。这些研究不仅有助于深入理解联邦学习中的后门攻击机制，还为构建更加安全可靠的联邦学习系统提供了重要的理论支撑和实践指导。

1.2 国内外研究现状

大数据时代，机器学习（Machine Learning）技术广泛应用于数据分析、模式识别，协助使用者提高效率，解决各领域的复杂问题。随着数据量的增大，模型变得更为复杂，机器学习有分布式的趋势，谷歌引入的联邦学习是一种分布式的机器学习模型：每个用户有一个本地训练数据集，这个数据集不上传到服务器。每个用户从服务器获取参数对数据进行训练，并且通过某种更新算法使得全局模型可以随着本地训练进行更新。由于联邦学习的高隐私性、高质量性和高独立性 [6]，它被广泛应用于政务服务、医疗系统、金融保险、物联网应用 [7] 等领域。随着联邦学习研究的持续推进，多个科研机构和公司纷纷发布了针对各类应用场景的联邦学习框架，如 FATE、TensorFlow Federated 和 PySyft 等 [8]，这些框架的提出为联邦学习领域的发展注入了新的活力。

Gu 等人 [9] 提出的后门攻击是深度神经网络面临的巨大安全威胁之一，这种数据中毒策略的核心在于有针对性地篡改训练数据集的特定子集，使机器学习模型在训练过程中被植入一种隐蔽的“后门”。这种后门使模型在面对嵌入特定触发器的测试集时，表现出异常的输出。Liu 等人 [10] 提出攻击者在线下载开放的预训练模型场景下实施后门攻击，攻击者不能访问原始训练集和验证集，但可以对预训练模型进行重训练，从而对面部识别、语音识别等模型进行后门攻击，不过该场景下的后门攻击比较容易防御。Yao 等人 [11] 提出的在迁移学习场景下的后门攻击，攻击者对预训练模型进行投毒训练，通过发布预训练模型实施后门攻击，借由用户下载来完成后门的植入。这种后门攻击相比之前的方法具有更强的隐蔽性和准确性。对于后门触发器的研究，Chen 等人 [12] 提出了不可见的后门攻击，通过使用不明显的后门图案来避免被发现，实现隐蔽的后门攻击。然而，过于隐蔽的后门图案不仅难以被人类发现，也会导致深度学习模型难以识别。为了解决这一问题，Liao 等人 [13] 研究了隐形触发器，设计了一种从视觉角度保证后门隐身的触发器，并以较低的注入率实施后门攻击。Li 等人 [14] 提出观察数据的特点，通过数据集自身的特点将后门触发器设置成隐藏在图案某特定部位的隐蔽后门触发器，而 Adi 等人 [15] 则通过将多个干净样本的正常标签进行组合的方式形成后门，通过样本叠加触发后门攻击。

由于联邦学习的分布式特性以及在传统的集中式机器学习中所提出的后门攻击方法直接依赖全局数据集的特点，在传统的集中式机器学习上提出的后门攻击方法难以直接应用于联邦学习的场景。因此，针对联邦学习场景，Bagdasaryan 等人 [16] 提出了集中式的后门攻击，在联邦学习过程中，全局模型不断聚合局部模型实现模型更新，当全局模型的准确率较高且趋于稳定时，攻击者将用中毒数据与

干净数据共同训练出来的中毒模型提交到聚合器，实现在全局模型中植入隐蔽的后门。因为联邦学习的分布性特征，各个参与者的数据和模型互不可见，后门的植入可以以更隐蔽的方式进行，因此后门攻击对联邦学习的威胁和危害是巨大的。然而在这种攻击方式下，由于联邦学习的聚合机制会削弱局部中毒模型对整体全局模型的影响，因此随着聚合轮次的增加，植入的后门可能被遗忘。此外，针对联邦学习的后门攻击还包括基于修改训练过程的后门攻击，这种攻击方法是指攻击者通过修改训练算法或修改训练过程中的模型参数来实现攻击，与仅对数据进行投毒的后门攻击相比，具有更强的隐蔽性。基于修改训练后模型的后门攻击，攻击者直接修改模型参数，如对模型进行缩放操作来实现攻击，这种直接修改模型参数来替换模型的攻击性能较强，但隐蔽性较差。基于边界情况的后门攻击，由于边界样本在训练和测试数据中出现频率较低，因此可以对边界样本数据进行投毒，从而植入后门。这种方法可以绕过如基于剪裁和添加噪声的防御方法、基于欧氏距离相似性的防御方法以及其他鲁棒聚合算法等多种防御方法且攻击寿命长，但只能对小概率出现的类别进行后门攻击。除了针对图像识别的联邦学习下的后门攻击外，林等人 [17] 设计了一种针对联邦学习的组合语义后门攻击，充分利用联邦学习的分布式特性，使该后门攻击方法在拜占庭聚合方法下也能有很好的攻击成功率和现实意义，且具有较强的隐蔽性，不易被检测。

随着对联邦学习场景下后门攻击研究的深入，后门攻击呈现攻击方法多样化、隐蔽性增强、攻击成功率增加、可跨设备协作的特点，攻击技术的不断改进要求联邦学习有更强的防御能力。后门攻击的防御方法包括基于差分隐私的防御方法 [18]，通过更新规范的阈值，添加高斯噪声来限制攻击的成功，针对后门攻击可能会产生的较大范数的更新，可以让服务器忽略范数超过某个阈值的更新或将其剪裁到某个范围之内，可确保每个模型更新的范数较小，来缓解恶意更新对服务器的影响。但是这种方法无法防御植入后门触发器的后门攻击，因此，Omid 等人 [19] 提出基于分组聚合的防御方法，通过服务器对客户端进行分组，分别对每个队列的训练模型进行聚合，在对每个队列分组聚合的结果进行检测，丢弃预期与良性模型检测结果不一致的模型后，再进行最终的聚合，更新全局模型。该方法可防御植入后门触发器的后门攻击且与安全聚合兼容，但要求恶意客户端数量要少。Mustafa 等人 [20] 提出通过使用调整参数学习率的聚合方法进行后门防御，该方法首先探讨了在联邦学习环境下成功实施后门攻击所必需的关键步骤，基于这些分析设计了一系列防御策略，旨在针对每个训练回合中的每一维度，根据客户端提供的更新符号信息，动态调整聚合过程中的学习率。这种方法不仅能够有效应对后门攻击，而且可以与现有的其他防御机制相结合，从而进一步增强系统的防御能力。然而，这种方法在提升防御效果的同时，也在一定程度上牺牲了部分数据隐私性。Hou 等人 [21] 提出了通过对投毒数据进行过滤的防御方法，服务器使用过滤器识别后门输入，将输入数据和模型的预测结果一起输入到后门过滤器中，将被后门过滤器视为可疑后门数据的输入数据及其预测向量保存到缓冲器中，使用去触发工具通过模糊触发区域来消除触发的影响，将去触发后的数据再次输入到模型中消除后门效应。Zhao 等人 [22] 提出通过提高模型稳定性来对后门攻击进行防御，在模型的训练过程中，引入参数的不确定性，用以提升模型的稳定性和泛化能力，进而增强了模型在面对对抗性样本时抵御潜在后门攻击的能力。为了实现这一目标，选择性地丢弃部分优化信息，从而在不牺牲过多性能的前提下，增加模型的泛化能力。通过这种方法，我们不仅能够增强模型的健壮性，还能使其在面对未知和复杂环境时，展现出更高的适应性。

联邦学习的后门攻击呈现出智能化、多样化的趋势，攻击者利用深度学习、强化学习等技术来优化攻击策略，提高攻击的成功率、持久性和隐蔽性，同时不断扩展应用场景，针对不同类型的数据和任务，设计更加精准和有效的攻击方法。为了应对日益严峻的后门攻击威胁，研究者们需要不断探索新的攻击方法和防御技术，并将多种防御技术融合使用，以提高联邦学习系统的安全性和鲁棒性。

1.3 思路方法

本文针对联邦学习中的集中式后门攻击没有充分利用联邦学习的分布式特性问题，对基于联邦学习的分布式后门攻击方法进行了研究，总结了联邦学习和后门攻击的相关知识，探究了集中式后门攻击的缺陷与不足，总结了分布式后门攻击的实施方法，我们通过模型替代的方法实施后门攻击，分别比较了单轮攻击和多轮攻击下分布式后门攻击和集中式后门攻击性能，改变分布式后门攻击触发因素探究触发因素对攻击成功率的影响，探究单触发模式后门攻击的成功率，使用不同的聚合方式探究了联邦学习的鲁棒性，思考分布式后门攻击的防御，为联邦学习的安全性保护提供思路，促进联邦学习在各个应用场景下的发展。

2 分布式后门攻击相关基础知识

2.1 深度学习

深度学习作为机器学习领域中新的研究方向，致力于研发更为先进的算法，以实现机器具备类似人类的分析学习能力，包含了对文字、图像和声音等多样化数据的准确识别与理解。深度学习通过构建多层的神经网络来模拟人脑复杂的神经元网络结构，并借助反向传播算法对模型进行训练，使模型学习样本数据的内在规律和表示层次。这些网络模型通常包括多个隐藏层，其中每个隐藏层都包含了大量的神经元节点。这些节点通过可调节的权重相互连接，模拟了生物神经网络中神经元间的信息传递过程。深度学习的核心思想在于通过多层次的特征抽象与转换，解决复杂的模式识别问题。在实际应用中，深度学习网络由众多紧密相连的神经元构成，它们逐层提取数据的特征信息，通过大规模数据集的训练，不断优化模型的性能，从而实现对复杂问题的精确处理。深度学习用于有监督、无监督、半监督、自监督、弱监督等的特征学习、表示、分类、回归和模式识别等 [23] 任务，广泛应用于数据挖掘、自然语言处理（NLP）[24]、语音识别 [25]、图像分类和识别、推荐和个性化技术等领域。

深度学习与传统的机器学习不同，主要表现在模型层次、特征提取和数据复杂度等方面。机器学习通常使用的是传统的线性模型或非线性模型，如决策树、支持向量机，而深度学习则使用多层神经网络，网络中的神经元之间存在大量的连接和权重，因此深度学习的模型结构比机器学习的模型结构更复杂，其包含更多的层次，特别是含有隐藏节点的层数。在机器学习中，通常需要人工进行特征提取，即人工在模型训练前从原始数据中提取出对模型训练有用的特征，并在模型训练和测试中对特征标签进行规定。而深度学习模型可以自动从原始数据中学习特征，提高了模型训练的效率，简化人工的操作。为了使深度学习的模型训练更有效，深度学习所用到的训练数据在质量和多样性上比机器学习有更高的要求，同时也在处理数据复杂的数据集上比机器学习更有优势。

深度学习分为监督学习、无监督学习和强化学习，涉及多种方法，包括卷积神经网络（CNN）、多层感知器（MLP）自编码神经网络和深度置信网络（DBN）等，在深度学习模型中，卷积神经网络的应用最为广泛和常见。神经网络是以神经元为基本单位，由相互连接的结点按层组织组成的，用来模拟大脑的某些机理与机制的结构，擅长处理图像识别、图像分类等计算机视觉任务。神经网络能够在无人工干预的情况下从数据中学习特征，进行自我学习和改进。卷积神经网络是一类包含卷积计算且具有深度结构的前馈神经网络，其模拟生物的视觉机制，能够按阶层结构对输入数据进行分类，具有表征学习能力。

卷积神经网络以局部连接、共享权重、池化和多层使用为基本思想，具有局部感知和参数共享的特性。图 2.1 是卷积神经网络的结构图，卷积神经网络的基本结构包括卷积层、激活函数、池化层、全连接层、输出层。其中卷积层是卷积神经网络最重要的一个层，由若干卷积单元组成，通过反向传播算法优化得到卷积单元的参数。卷积层的作用是对输入数据进行卷积操作，进行局部关联和窗口滑动，通过迭代提取输入数据的不同特征。卷积神经网络的每个卷积层都包含多个滤波器或卷积核，激活函数就是决定是否向下一个层传递信息的关键组件，最常见的激活函数包括 ReLU、Sigmoid 和 Tanh 等。池化层的作用是对数据进行降维处理，压缩数据和参数的量且具有特征不变性，他将在卷积层之后得到的维度很大的特征切成几个区域，取其最大值或平均值，以得到新的、维度较小的特征，防止过拟合。全连接层作为分类任务中神经网络的最后一层，将数据矩阵进行全连接，然后按照分类数量输出数据，把所有局部特征结合变成全局特征，用来计算最后每一类的得分。

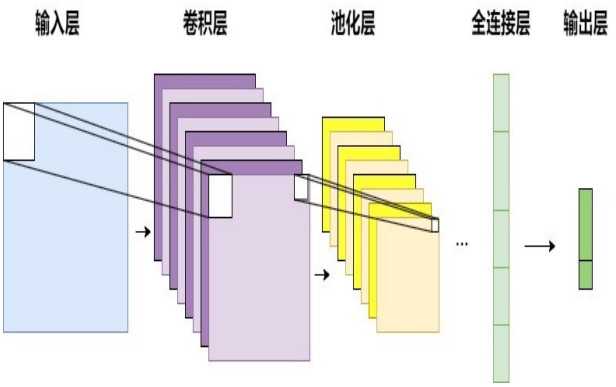


图 2.1 卷积神经网络结构图

卷积神经网络的训练过程分为前向传播阶段和反向传播阶段，如图 2.2 所示，在前向传播阶段选取训练样本输入网络中，随机初始化权值，从输入层对数据进行一层一层的特征提取和转换，在输出层得到输出结果；在反向传播阶段将输出结果与理想结果对比，计算全局误差将得到的误差反向传递给不同层的神经元，通过“迭代法”调整权值和偏重，来寻找全局性最优的结果。

2.2 联邦学习

随着大数据和人工智能技术的飞速发展，数据

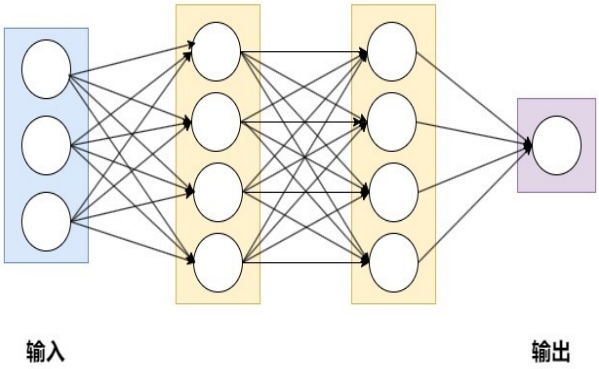


图 2.2　全连接神经网络原理示意图

已经成为驱动技术进步和创新的关键因素，且数据有分散在各个企业或机构中保存的特性。由于数据所有权、隐私安全、行业竞争等因素，在实际应用中数据的收集、存储和共享面临着诸多挑战，直接共享原始数据往往难以实现，同时，随着社会对数据隐私的重视和有关数据安全法规的出台，复杂而量多的数据需要更复杂的训练模型，和大量的计算资源需求。因此，传统的集中式训练模型的方式已经难以满足大量分散的数据的处理需求。为了解决这种问题，联邦学习应运而生，其目的在于在保障数据隐私和安全的前提下，实现多个参与方之间的协同训练，共同提升模型的性能和效果。通过让各个参与方在本地设备上训练模型，并将模型的更新参数发送到中央服务器进行聚合，联邦学习可以充分利用各个参与方的数据资源，同时保护他们的数据隐私和安全性。

联邦学习作为一个流行的机器学习框架，具有分布式的特征，允许客户以分散的方式训练机器学习模型，而无需共享任何私有数据集。联邦学习的核心理念是“数据不动模型动，数据可用不可见”。具体来说，它强调在训练过程中不直接共享原始数据，而是通过共享模型的更新参数来实现协同训练。这样可以在保护数据隐私和安全性的同时，实现多个参与方之间的数据共享和协同学习。除此之外，联邦学习还强调“分布式优化”的思想。即在联邦学习的训练过程中，各个参与方在本地设备上使用自己的数据进行模型训练，并将训练得到的模型参数发送给中央服务器。中央服务器通过对这些参数进行聚合和更新，得到一个全局模型，并将其返回给各个参与方进行下一轮的训练。通过这种方式，联邦学习可以实现分布式环境下的高效优化和模型训练。

联邦学习、集中式学习（Centralized Learning）和分布式学习（Distributed Learning）是机器学习领域中不同的学习方式，联邦学习可以看做是集中式学习和分布式学习方法的结合，如图 2.3 所示，集中式学习是将整个系统视为一个整体，需要所有参与者进行全局通信来共享数据和模型信息，其采用单智能体强化学习算法，将数据集中在一起进行集中式训练。集中式学习的可扩展性较差，参与者数量有限，数据隐私保护较弱，有数据泄露的风险，因此，集中式学习适用于数据量较小、计算资源集中、无需过多考虑数据隐私保护的场景。分布式学习是将数据集进行分割，然后将每个数据块发送到不同的设备上进行本地训练。分布式学习相比于集中式学习来说可以降低通信开销，更好的数据隐私保护，但是由于分布式学习的训练都是在本地的，因此每个本地模型之间的差异性较大，需要更多的计算资源和迭代次数来达到相同的训练效果，学习效率不高。分布式学习适用于数据量较大、需要降低通信开销和保护数据隐私的场景。为了弥补集中式学习和分布式学习的缺点，联邦学习将两者的思想进行结合，使得参与各方在不共享本地数据的前提下，进行多方的协同训练。联邦学习拥有数据绝对掌握权，每一个参与方数据都不离开本地，模型信息在各参与方之间以加密的形式传输，保证不能由模型推测出原始数据，且联邦学习由中央服务器对训练过程进行聚合统一，学习效率高。但联邦学习仍存在参与方不稳定、通信代价高等缺点，适用于需要保护数据隐私、数据分散在多个不同地方且难以集中的场景。

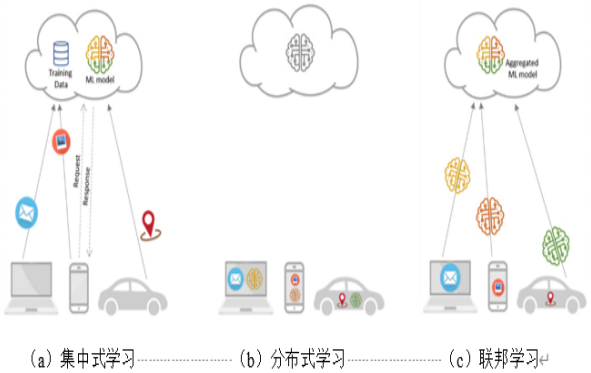


图 2.3　集中式学习、分布式学习和联邦学习的比较

联邦学习利用分布式学习技术，将深度神经网络的训练分布到参与者的本地模型上，中央服务器负责在每一轮迭代中选择特定的参与者子集，这些客户端将其训练得到的本地模型的结果与中央服务器进行共享，中央服务器将局部模型聚合成全局模型后，将更新后的全局模型分发给各个客户端，各客户端基于接收到的全局模型来更新它们各自的本地模型，从而进入下一轮迭代，这一循环过程将持

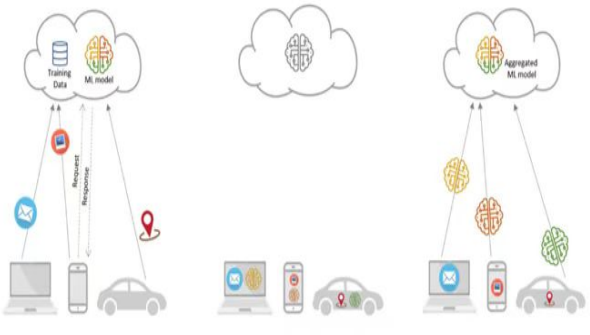
续进行，直至全局模型收敛标准。图 2.4 是联邦学习示意图，该图中三个参与者分别训练自己的局部模型，将局部模型发送到聚合器（中央服务器）进行聚合，聚合后聚合器将更新的全局模型再发送回参与者进行下一轮的局部训练，该过程不断迭代直至模型收敛。

假设有 n 个参与者 $\{C_1,\ldots,C_n\}$，每个参与者对应拥有本地数据 $D=\{D_1,\ldots,D_n\}$，传统的机器学习方法是由一个服务器使用所有数据 D 训练模型，该模型的准确率精度值表示为 V_{SUM}，而联邦学习是所有参与者 C_i 使用本地数据 D_i 进行数据训练后由中央服务器将本地训练模型聚合生成全局模型的过程，全局模型的准确率精度值为 V_{FED}，该过程中参与者的本地数据都不会与其他参与者共享。联邦学习的精度值 V_{FED} 应非常接近于 V_{SUM}，我们用 δ 来表示联邦学习的精度值损失，δ 为一个非负实数，其满足：$\delta=|V_{\mathrm{SUM}}-V_{\mathrm{FED}}|$。$|V_{\mathrm{FED}}-V_{\mathrm{SUM}}|<\delta$（式 2-1）

每个轮次中，中央服务器向所选择的参与者发送当前的聚合模型 G^t，每个 C_i 使用其本地数据集 D_i 更新其本地模型参数 w_i^t，其中 t 表示当前迭代轮。参与者本地训练将模型更新为新的本地模型 $L_i^{(t+1)}$，并将差值 $L_i^{(t+1)}-G^t$ 发送回中央服务器，中央服务器对接收到的模型更新进行平均，得到新的全局模型：

$$G^{(t+1)}=G^t+\frac{\eta}{n}\sum_{i=1}^{m}\left(L_i^{(t+1)}-G^t\right)$$

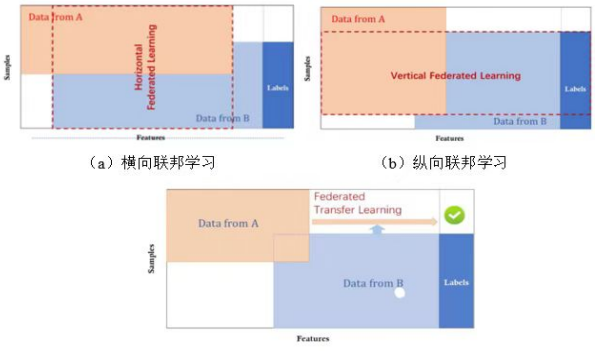
（式 2-2）其中 n 为联邦学习总参与者数量，m 为每轮训练的参与者数量，η 为全局学习率，用来控制每轮更新的全局模型的比例，若 $\eta=n/m$，则表示模型完全被局部模型的平均值所代替。



（a）集中式学习　（b）分布式学习　（c）联邦学习

图 2.4　联邦学习示意图

根据数据特征和样本在客户端之间的分布情况，联邦学习被分为横向联邦学习、纵向联邦学习和联邦迁移学习。图 2.5 展示了横向联邦学习、纵向联邦学习和联邦迁移学习在数据样本和数据特征上的重叠特点。横向联邦学习是基于样本的联邦学习，适用于参与方拥有的数据样本不同但数据有重叠的数据特征。例如，两个地区的银行拥有各自区域不同的用户集，用户交集非常小。但是，由于他们的业务非常相似，因此用户的特征空间是相同的。纵向联邦学习也叫按特征划分的联邦学习，适用于参与方训练数据在数据样本上有重叠，但在数据特征上有所不同的情况。例如，同一地区的银行和电子商务公司，他们的用户集可能包含该地区的大多数居民，因此他们的用户空间交集很大。但是，由于银行和电子商务业务的不同，所以他们记录了用户不同方面的数据，特征空间有很大不同。联邦迁移学习适用于两个数据集重叠也很少，特征空间也不同的情况。例如，A 地区的一家银行和 B 地区的一家电子商务公司，一方面，由于地区的限制，两个机构的用户有很小的交集，因此数据集重叠小，另一方面，由于两个机构业务不同，因此双方的特征空间只有小部分重叠。在当前的场景下，联邦迁移学习技术被有效应用于基于有限的公共样本集，学习两个特征空间之间的共同表示，这一共同表示被进一步应用于仅包含单边特征的样本的预测任务中，通过这种方法为联邦学习框架下的整个样本和特征空间提供了解决方案。



（a）横向联邦学习　（b）纵向联邦学习

（c）联邦迁移学习

图 2.5　联邦学习的分类

2.3　后门攻击

神经网络后门攻击是指在神经网络的训练阶段，攻击者利用神经网络对非鲁棒特征的过度学习能力植入后门，使该神经网络的训练结果在良性数据上有正确输出且无异常表现，而对于有特定后门的输入数据，神经网络会错误的将其分类为攻击者选择的目标标签而不影响全局准确率。后门攻击允许攻击者在系统中执行恶意操作而这些恶意操作对操作系统的用户不可见，从而增加了后门攻击的检测和防御难度。在后门攻击场景中，触发器是用于生成中毒样本的特定图案，通常以像素图案的方式出现，目的在于激活隐藏后门。良性样本是那些未

经任何篡改或污染的原始数据样本，而中毒样本是被植入触发器的数据样本，在训练阶段，中毒样本被注入训练数据集中，用于在模型里嵌入后门，在测试阶段，任意被插入触发器的中毒样本都能触发后门机制，最终被错误地分类到攻击者设定的目标类别中，实现后门攻击。

图 2.6 展示了后门攻击的攻击原理和实现方法，以该图为例，图中 (a) 是后门触发器的示意图，后门触发器是位于图片数据右下角的白色像素方块，其作用是将数据的原始标签修改为目标标签，设置目标标签为黄色，即我们期望被植入后门触发器的数据输出的分类结果为目标标签黄色。在训练阶段，我们使用良性数据以及被后门触发器污染的后门数据组成的数据集进行训练，生成一个后门模型。在测试阶段，用该后门模型分别对带有后门触发器数据和干净的数据进行结果预测，植入后门触发器的数据分类结果为目标标签黄色，而干净数据分类结果为其正确标签，即该后门模型在良性数据上预测结果准确而在后门数据上使其预测结果为目标标签，后门攻击成功。

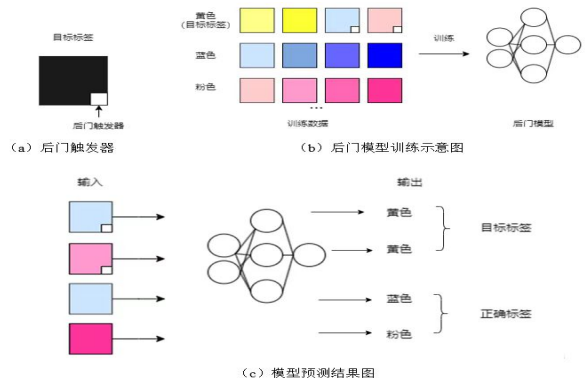


图 2.6　后门攻击原理示意图

后门攻击是由攻击者产生一个恶意的后门模型，攻击者的目标有两个，一个是后门模型在良性数据上的准确率要与良性模型一致，即后门攻击不能被用户察觉。第二个是对于攻击者选择的特定标签的输入，即包含后门触发器的输入，后门模型输出的预测结果要与经过真实训练的模型的预测结果不同。因此后门攻击的训练目标函数通常涉及两部分，一部分是针对干净训练数据的经验损失项，另一部分是针对后门训练数据的经验损失项。后门攻击训练目标函数如下：

$$L_{\text{total}} = \alpha \cdot L_{\text{clean}} + \beta \cdot L_{\text{backdoor}} \qquad \text{(式 2-3)}$$

其中 L_{total} 是训练模型的总体目标函数，L_{clean} 是针对正常训练数据的经验损失项，通常使用交叉熵损失或其他适当的损失函数，L_{backdoor} 是针对后门训练数据的经验损失项，其目标是在模型中加入后门，具体形式取决于攻击者的目标和所使用的后门类型。α 和 β 分别是正常训练数据的经验损失项和后门训练数据的经验损失项的权重参数，用于控制两部分损失在总体目标函数中的相对重要性，可以根据攻击者的需求进行调整。

后门攻击可能出现在三种典型场景中：用户为降低训练成本采用第三方数据集进行训练、基于第三方平台 (如云计算平台) 训练神经网路、直接使用第三方模型进行训练，三种场景分别通过 Internet 向用户提供有毒数据集、在实际训练过程中对数据集和训练计划进行修改以及通过应用程序编程接口 (API) 和 Internet 提供经过训练的受感染的神经网络模型的方式实现后门攻击，以上三种场景使用户失去对训练阶段的控制权或知情权，进一步扩大训练模型的安全风险。从可见性的角度，可以将后门攻击分为两类，可见后门攻击和不可见后门攻击，可见后门攻击是指攻击者在训练数据集中插入明显的、容易被检测到的触发器，不可见后门攻击是指攻击者插入不明显或难以被检测到的触发器，这种后门更加隐蔽。可见后门攻击包括 Gu 等人 [9] 提出的 BadNets 攻击是在神经网络训练阶段插入后门实施攻击，攻击者通过在训练数据中添加特定的标签和触发器进行后门植入。不可见的后门攻击包括 Chen 等人 [12] 提出的通过使用不明显的后门图案来实现隐蔽的后门攻击、Liao 等人 [13] 提出的使用隐形触发器，并使用较低的注入率实施后门攻击、Li 等人 [14] 提出的将后门触发器隐藏在图案某特定部位的后门攻击和 Adi 等人 [15] 提出的将多个干净样本的正常标签进行组合的方式形成后门进行后门攻击等。

2.4　联邦学习中的后门攻击

由于联邦学习训练的分布性，其训练过程容易受到攻击，其中，后门攻击是联邦学习安全性问题的一大威胁。在联邦学习中，可能存在模型鲁棒性和隐私保护的安全性问题，其中，模型鲁棒性问题包括拜占庭攻击和后门攻击，这两种攻击方式都是由恶意客户端攻击服务器，针对训练数据或针对局部模型投毒实现的，区别是拜占庭攻击是无目标攻击，会影响全局模型性能而后门攻击的攻击目的为通过攻击影响目标子任务的性能。隐私保护问题包括推理攻击，是由恶意服务器或恶意客户端攻击客户端，针对客户端参数或全局模型参数推理进行攻击，目的是获取客户端的信息。

在联邦学习场景中，后门攻击的实施方式主要包括针对训练数据的攻击和针对局部模型的攻击。在探讨针对训练数据的攻击时，我们假设攻击者具备的是黑盒攻击能力，这意味着攻击者仅能在有限

的范围内修改恶意客户端的局部训练数据，而无法干预或修改其训练流程和最终生成的模型。而对于针对局部模型的攻击，我们则假设攻击者拥有更为强大的白盒攻击能力,在此情境下,攻击者不仅能够完全控制恶意客户端的局部训练数据和训练流程,还能直接对训练后得到的模型进行修改和调整。

在联邦学习中，后门攻击可以发生在模型聚合的任何一个过程中，即后门攻击可以出现在模型聚合前、聚合中和聚合后。在模型聚合前，攻击者主要通过改变训练数据或局部模型来实施攻击，可以进行数据投毒或局部模型投毒。数据投毒是攻击者通过向训练数据集中注入带有特定触发器的样本,使模型在训练过程中学习到这些后门触发器与特定输出之间的关联，以便在模型部署后，对包含这些触发器的输入数据做出异常的输出。局部模型投毒是攻击者对一部分参与训练的客户端或节点进行控制，在这些客户端或节点上训练出带有后门的局部模型，并在聚合过程中将这些模型的影响传播到全局模型中。在聚合过程中，攻击者可以通过修改聚合算法或参数来实施攻击。攻击者可以通过调整不同局部模型的权重，使得带有后门的局部模型在聚合过程中占据更大的比重，从而增强后门对全局模型的影响。也可以通过调整聚合过程中的学习率等超参数，使模型更容易学习到后门信息。在模型聚合后，攻击者主要通过对全局模型进行后处理或修改来实施攻击。然而在联邦学习中由于攻击者无法直接访问或修改全局模型，因此在聚合后实施后门攻击的难度较大。

针对后门攻击，联邦学习模型也有相应的防御手段，其防御可以在模型训练聚合任何一个时间实施，包括训练时、聚合前、聚合中和聚合后。训练时的防御主要是检查训练数据或修改训练过程，而其他和聚合过程相关的防御手段则是检查或修改客户端提交的模型结果。聚合前的主要防御方法主要有基于差分隐私的防御方法、基于降维的防御方法和基于修改协议过程的防御方法，基于差分隐私的防御方法通过裁减更新、添加高斯噪声来实现；基于降维的防御方法通过提取高维模型的特征，在低维特征空间中进行判别来防御后门攻击；基于修改协议过程的防御方法通过修改协议过程检测并剔除恶意更新。聚合中的防御方法主要有基于相似性的防御方法和基于统计的防御方法，基于相似性的防御方法利用欧式距离、余弦相似度等计算模型的相似性并根据相似性调整局部模型的权重或学习率来实现防御；基于统计的防御方法根据统计特性选择具有代表性的客户端更新并估计真实更新的中心。聚合中的主要防御方法主要有基于联邦遗忘的防御方法和基于全局模型性能的防御方法，基于联邦遗忘的防御方法是在联邦训练结束后，从全局模型中彻底删除某个客户端的贡献以完全消除恶意客户端的影响；基于全局模型性能的防御方法是利用客户端的本地数据集测试全局模型来验证全局模型是否具有后门。

3 针对联邦学习的分布式后门攻击

3.1 集中式后门攻击

传统的后门攻击通常是集中式后门攻击，通过破坏训练数据来改变模型的输出结果。这些攻击方法仅通过数据中毒或向固定模型插入后门组件来改变模型在攻击者选定的数据上的行为。然而,由于联邦学习是多个局部模型聚合在一起形成全局模型,聚合过程会抵消后门模型的效果，使得攻击者的模型容易被遗忘。因此，传统的集中式后门攻击对联邦学习的效果不佳。联邦学习中，攻击者能够完全控制一个或多个参与者，通过控制参与者的本地训练数据、控制局部训练过程和学习次数、学习率等超参数、在提交聚合之前修改结果模型的权重等操作来自适应地改变局部训练。攻击者的目标是使联邦学习产生一个全局模型，该模型在其主任务和攻击者选择的后门子任务上都能达到高精度，并且在攻击后的多个回合中保持后门子任务的高精度。后门攻击的目的是使训练模型在任何嵌入后门触发器输入数据上预测结果为攻击者选定的目标标签 τ。联邦学习中攻击者 i 在第 t 轮与局部数据集 D_i 和目标标签 τ 的对抗目标为:

$$w_i^* = \arg\max_{w_i} \Bigg(\sum_{j \in S_{\text{poi}}^i} P\big[G^{(t+1)}(R(x_j^i, \phi_i^*)) = \tau \mid \gamma; I\big] + \sum_{j \in S_{\text{cln}}^i} P\big[G^{(t+1)}(x_j^i) = y_j^i\big] \Bigg), \quad \forall i \in [M] \tag{式 3-1}$$

其中中毒数据 S_{poi}^i 和干净数据 S_{cln}^i 满足 $S_{\text{poi}}^i \cap S_{\text{cln}}^i = \emptyset$ 和 $S_{\text{poi}}^i \cup S_{\text{cln}}^i = D_i$。函数 R 使用一组参数 ϕ 将任意类中的干净数据转换为具有攻击者选择的触发图案的后门数据。对于图像数据来说参数 ϕ 可以被分解为触发位置 TL、触发尺寸 TS 和触发间隙 TG, 即 $\phi = \{TS, TG, TL\}$), 攻击者可以通过改变参数 ϕ 相关因素的值来设计自己的触发图案，并选择一个最优的数据中毒比 r 来得到一个更好的模型参数 w_i^*，全局模型可以通过参数 w_i^* 为后门数据分配最高概率的目标标签 τ，也可以为良性数据分配全局真实标签 y_j^i。

在联邦学习中，采用模型替换的方法实施后门攻击比较常见。在这种方法中，攻击者试图用恶意模型 X 代替新的全局模型 $G^{(t+1)}$，后门攻击步骤

如下：

- 迭代训练：攻击者对模型进行迭代训练，如果现模型损失小于后门任务的最大损失，则停止迭代，否则，则用数据集中的数据替代批数据中的数据。
- 更新模型：用当前模型减去攻击者的学习率与当前批数据在该模型下的损失梯度的乘积来更新模型。
- 调整学习率：每次迭代结束时，若此时攻击者的学习率应该降低（提高持久化），则用攻击者学习率除以学习率的下降来调整攻击者的学习率。
- 模型放大：迭代完成后，在提交模型前按比例放大模型。

在模型训练时，每个局部模型可能与当前的全局模型相差甚远，但随着全局模型的收敛，这些偏差开始被抵消，通过模型替代算法，可以求出攻击者提交的局部模型为：

$$\tilde{L}_m^{(t+1)} = \frac{n}{\eta}X - \left(\frac{n}{\eta} - 1\right)G^t - \sum_{i=1}^{m-1}(L_i^{(t+1)} - G^t) \approx \frac{n}{\eta}(X - G^t) + G^t \text{(式 3-2)} \tag{1}$$

这种攻击通过比例因子 $\gamma = n/\eta$ 将后门模型 X 的权重放大，以确保后门在聚合平均中保留下来，并且全局模型被 X 取代。在这种后门攻击方法下，损失函数被定义为：

$$L_{\text{model}} = \alpha \cdot L_{\text{class}} + (1 - \alpha) \cdot L_{\text{ano}} \tag{式 3-3}$$

其中，L_{class} 捕获了主任务和后门任务的准确性，L_{ano} 为异常检测项，可以用于任何类型的异常检测，这个算法通过添加异常检测项 L_{ano} 来修改目标（损失）函数，以规避异常检测，通过超参数 α 用来控制回避异常检测的重要性。

由于集中式后门攻击没有充分利用联邦学习的分布式特性，因此，集中式后门攻击存在一定的缺陷。集中式后门攻击在实施攻击过程中通常只有一个攻击者来植入后门，如果该攻击者被识别或阻止，整个攻击就会失效，攻击的可持续性受到限制。集中式后门攻击通常将相同的全局触发模式嵌入到所有对抗方的数据中，这种触发方式导致一旦攻击者使用的触发模式被检测到或分析出来，防御者就能迅速识别并防御所有的攻击实例，使得模型对后门攻击的可检测性和可防御性大大提高。同时，由于集中式后门攻击触发模式是全局统一的，攻击行为在多个参与者之间表现出一致性，这使得攻击行为容易被防御系统识别，后门攻击缺乏隐蔽性。

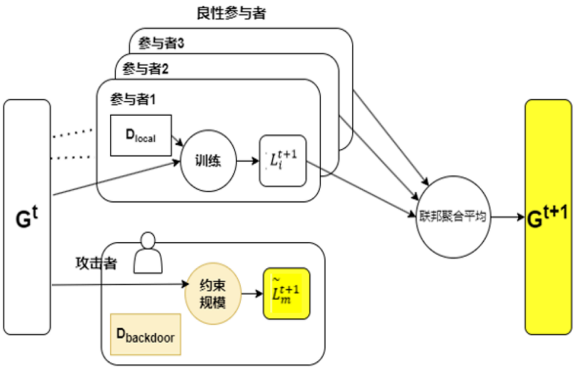


图 3.1　联邦学习集中式后门攻击的训练原理图

3.2　分布式后门攻击

由于集中式后门攻击可能导致攻击失效或攻击性能不好，我们考虑提出一种充分利用联邦学习的分布式特性的分布式后门攻击方法，使后门攻击更适应局部模型，将攻击分给不同的攻击者以保证攻击的有效性，同时通过细分攻击过程提高攻击成功率。

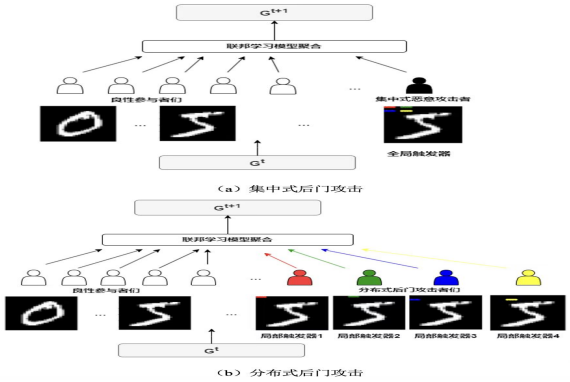
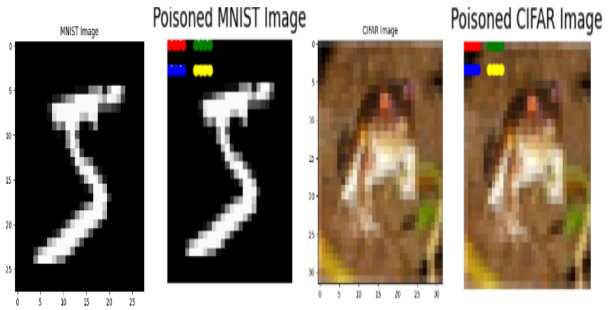


图 3.2　集中式后门攻击与分布式后门攻击比较

联邦学习场景下传统的集中式后门攻击 [15] 采用将相同的全局触发图案嵌入到所有攻击方的方法实施后门攻击，而我们的方法使用与集中式攻击相同的全局触发器，但将其分解为本地触发器分别嵌入到不同的攻击方，该分布式后门攻击方法充分利用联邦学习的分布式特性，其攻击性能明显优于集中式后门攻击，因为它的本地触发器更隐蔽，因此更容易绕过鲁棒聚合规则，攻击更隐蔽、更有效。如图 3.2 所示，集中式后门攻击采用全局触发器，分

布式后门攻击采用局部触发器。分布式后门攻击具有隐蔽性、高效性和对抗性等特性，同时对联邦学习的安全问题研究很有意义。在分布式后门攻击中，由于每个恶意客户端使用的局部触发器都不相同，因此这种攻击难以被检测和防御，特别是对于传统的防御机制来说，其难以识别出这种分散且多样化的攻击模式，攻击表现为隐蔽性。分布式后门攻击利用了分布式学习的特性，使得攻击者能够在不暴露原始数据的情况下，通过修改本地模型参数来实现对共享模型的后门攻击这种攻击方式具有更高的灵活性和效率，使攻击具有高效性。分布式后门攻击作为一种对抗性攻击方式，能够绕过传统的安全控制机制，实现对系统的非法访问和控制，该攻击具有对抗性。

在我们提出的分布式后门攻击中，我们向数据集中插入四个局部触发器构成全局触发器，图 3.3 是在 MNIST 和 CIFAR10 中插入后门触发器的示意图，为便于观察，我们分别用四个颜色标记了四个局部触发器，每个颜色标记的矩形像素图案为单个攻击者植入的后门，在实际训练中，后门触发器为四个白色矩形像素图案，设置四个局部触发器的大小、间隔相同。



(a) MNIST 触发器　　(b) CIFAR10 触发器

图 3.3　干净数据与带触发器数据

在联邦学习的分布式后门攻击框架下，有许多与局部触发器相关的新的触发因素：

触发大小 TS：规定了局部触发器的像素宽度。

触发间隔 TG：表示左、右、上、下两个局部触发的距离。

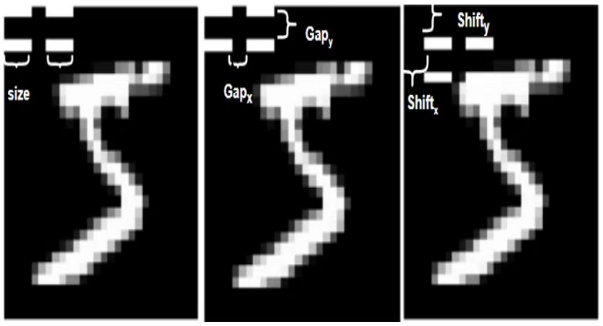
触发位置 TL：是触发模式与左上角像素的偏移量。

尺度 ：攻击者用来扩大恶意模型权重。

中毒比 r：用来控制每个训练批次添加的后门样本的比例

中毒间隔 I：两个中毒步骤之间的回合间隔。

我们使用了触发大小相同的矩形触发器作为局部触发器，并且两个局部触发器的触发间隔相同，在本文的实验部分，我们通过改变某些触发因素来探究触发因素对分布式后门攻击效果的影响。



(a) 触发大小　　(b) 触发间隔　　(c) 触发位置

图 3.4　触发因素

训练过程中共有 100 的参与者，每轮选择 10 个参与者，其中一部分是攻击者控制的恶意参与者，其余是良性参与者，恶意参与者为我们事先确定好的，而良性参与者从去掉攻击者后所有参与者中随机选出。我们的攻击设置了 4 个参与者，每个参与者在固定轮次使用一个局部触发器进行攻击。数据在参与者之间分布式存储，每个参与者都有自己的本地数据集，攻击者只能修改其控制的客户端的数据集。参与者之间通过中心服务器进行通信和模型聚合。攻击者能够将其修改的模型参数或梯度信息发送到中心服务器，并参与全局模型的聚合过程。系统使用的模型结构是已知的，并且攻击者可以访问和修改其控制的客户端上的模型参数和结构。

DBA 充分利用了联邦学习的分布式学习和本地数据不透明性，假设有 M 个攻击者和 M 个局部触发器，每个 DBA 攻击者 m_i 都独立地使用其局部触发器对其本地模型实施后门攻击。在这种机制中，将一个全局的攻击问题分解为 M 个分布式的子攻击问题，分布式后门攻击的攻击目标为：

$$w_i^* = \arg\max_{w_i} \left(\sum_{j \in S^i} P\left[G^{(t+1)}(R(x_j^i, \phi_i^*)) = \tau \mid \gamma; I\right] + \sum_{j \in S^i} P\left[G^{(t+1)}(x_j^i) = y_j^i\right] \right), \quad \forall i \in [M] \tag{2}$$

式 (3-4)

其中，τ 是后门攻击者选定的目标标签，所有中毒数据在模型中的输出结果要与 τ 值一致，$G^{(t+1)}$ 表示第 $t+1$ 轮的全局模型，S^i_{poi} 为中毒数据集，S^i_{cln} 为干净数据集，$\phi_i^* = \{\emptyset, O(i)\}$ 是攻击者 m_i 局部触发模式的几何分解策略。与集中式后门攻击不同的是，DBA 攻击者会在间隔 I 的毒害下毒，并在提交给聚合器之前使用比例因子 γ 来操纵他们的更新。

用模型替代法实施后门攻击的伪代码如下：

表 1　方法与参数说明

名称	说明
方法：	
$L_{\text{ano}}(X)$	模型 X 的异常检测，用来修改损失函数
$\text{replace}(c, b, D)$	数据替代，将 data batch b 中的 c 项替换为数据集 D 中的项
$\text{add}(\text{image}, \text{index})$	向数据集的图片中注入相应的局部后门触发器
参数：	
Lr_{adv}	攻击者的学习率
α	控制逃避异常检测的重要性
Step_sched	学习率应该降低的轮次
Step_rate	学习率的下降
c	需要替换的良性因子的数量
γ	比例因子
E_{adv}	攻击者的局部攻击轮次
ϵ	后门任务的最大损失
index	局部后门触发器的编号

算法 1．　模型替代攻击.

输入：G^t 上一轮训练的模型
输出：$\tilde{L}_m^{(t+1)}$ 训练好的看起来不异常的局部模型

//初始化供给模型 X 和损失函数 ℓ：
$X \leftarrow G^t$
$\ell \leftarrow \alpha L_{\text{class}} + (1-\alpha) L_{\text{ano}}$
for epoch $e \in E_{\text{adv}}$ do
  if $L_{\text{class}}(X, D_{\text{backdoor}}) < \epsilon$ then
    // Early stop, if model converges
    break
  end if
  for batch $b \in D_{\text{local}}$ do
    $b \leftarrow \text{replace}(c, b, D_{\text{backdoor}})$
    $X \leftarrow X - \text{Lr}_{\text{adv}} \cdot \nabla \ell(X, b)$
  end for
  if epoch $e \in$ step_sched then
    $\text{Lr}_{\text{adv}} \leftarrow \text{Lr}_{\text{adv}}/\text{step\_rate}$
  end if
end for
// Scale up the model before submission.
$\tilde{L}_m^{(t+1)} \leftarrow \gamma(X - G^t) + G^t$
return $\tilde{L}_m^{(t+1)}$

4　实验

4.1　实验准备

4.1.1　评估准则

实验过程中我们使用攻击准确率（Attack Success Rate，ASR）来评估模型性能，攻击准确率是指攻击者在目标模型上成功触发后门并达到其预期目标（如错误分类）的比例。要进行攻击准确率的评定，首先要创建一个包含后门触发器的测试集，用于评估模型在触发后门时的性能。然后使用带有后门触发器的测试集对目标模型进行测试，并计算模型错误分类的比例，这个比例就是攻击准确率。

除攻击准确率以外，考虑到后门攻击在良性样本上的表现以及后门攻击的持久性和模型鲁棒性，良性准确率（Benign Accuracy，BA）、存活时间和鲁棒性也是重要评估标准。良性准确率是模型在未经任何修改的原始数据集上的准确率，即模型在“良性”或正常情况下的性能。存活时间是指后门在目标模型中持续有效的时间长度。由于分布式学习系统通常是持续更新和迭代的，因此后门可能在一段时间后失效。评估存活时间需要考虑迭代次数、更新策略以及防御措施。鲁棒性是用来评估后门在不同条件下的稳定性，我们通过改变聚合算法的方式来评估模型的鲁棒性。

4.1.2　数据集

为了验证分布式后门攻击比集中式后门攻击更有效、更持久，我们选了两个数据集 MNIST 和 CIFAR10-10 来对 DBA 进行评估。MNIST 数据集是一个手写数字图像数据集，包含了从 0 到 9 的手写数字图像。训练集包含 60,000 个样本，测试集包含 10,000 个样本，每个图像都有一个对应的标签（0-9），表示图像中的手写数字。CIFAR10 数据集包含了一系列彩色图像，用于图像分类任务。我们所用的 CIFAR10 含 50,000 个样本，测试集包含 10,000 个样本，共有 10 个类别，包括飞机、汽车、鸟、猫、鹿、狗、青蛙、马、船和卡车。

对训练参数的设置，我们使用了 SGD 随机梯度下降算法，进行 E 个本地轮次，局部学习率为 l_r，批大小（batch size）为 64。所有参与者共同训练一个共享的全局模型，每轮选取 10 个参与者进行聚合。MNIST 数据集包含 0 到 9 的 10 个类别，每个数字图像都是 28×28 像素，因此特征数量是 784（28×28）。CIFAR10 是一个包含 10 个不同类别的小型图像数据集，每个图像都是 32×32 像素，并且有 RGB 三个颜色通道，因此特征数量是 1024（32×32×3）。MNIST 使用了包含两个卷积层（conv）和两个全连接层（fc）的神经网络模型，而 CIFAR10 使用了轻量级的 ResNet-18 模型。在投毒比例中，64 是指批次大小，即在模型训练过程中，一次迭代中用于更新模型权重的样本数量。在 MNIST 中一个批次中中毒的样本有 20 个，CIFAR10 中中毒的样本有 5 个。

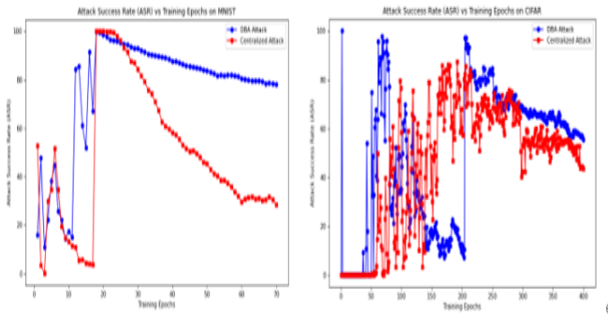
4.2　分布式后门攻击与集中式后门攻击

4.2.1　单轮攻击场景

我们使用模型替换的方式对联邦学习实施后门攻击，即每个 DBA 攻击者或集中式攻击者只需要在一个轮次中进行后门植入就可以成功向模型中植入后门触发器。为了实现一次植入后门触发器，攻击者在他们的恶意更新中执行缩放

操作使后门模型更新中的贡献超过其他良性模型，并确保后门在聚合过程中存活下来。对 MNIST 数据集，在分布式后门攻击中，我们选择第 12、14、16、18 轮分别植入 4 个局部后门触发器，在集中式后门攻击的第 18 轮植入 4 个局部触发器组合成的全局触发器，总共选取 10 个参与者，其中良性参与者随机抽取；对 CIFAR10 数据集，DBA 中我们选择第 203、205、207、209 轮分别植入 4 个局部后门触发器，在集中式后门攻击的第 209 轮植入全局触发器，共选取 10 个参与者。我们的实验分别使用全局触发器和局部触发器对后门攻击的攻击成功率进行测试，使用相同的全局触发器来评估分布式后门攻击和集中式后门攻击的不同。为了保证公平，我们确保 DBA 的后门触发器像素总数与集中式后门攻击的后门触发器像素总数接近或小于集中式攻击，MNIST 和 CIFAR10 中 DBA 全局触发器像素总数与集中式全局触发器像素总数的比值分别为 0.964 和 0.990。为了避免攻击成功率的测试结果受到原始标签的影响，我们在加载数据时将以与攻击者选定的后门标签为真实标签的数据去除，在全局模型开始收敛时开始攻击，数据集的全局学习率 为 0.1。

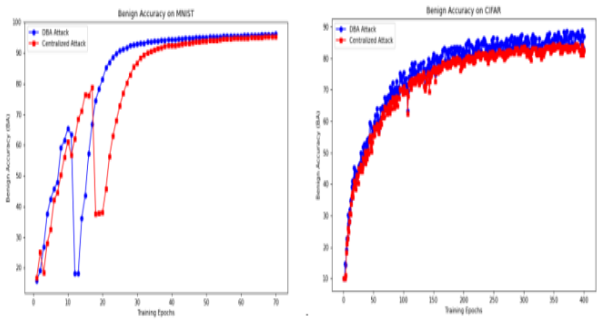
图 4.1 是在单轮攻击场景下集中式后门攻击与分布式后门攻击成功率的比较，其中。红色折线为集中式后门攻击，蓝色折线为分布式后门攻击。在单轮攻击场景中，DBA 和集中式攻击在以 = 100 为尺度因子的所有数据集上注入一个完整的后门后，攻击成功率都很高，这表明单轮攻击是有效的。后门触发器成功注入后，在连续的轮次中，被聚合到全局模型的后门模型被良性更新削弱，攻击成功率呈下降趋势，且集中式攻击在全局触发器下的攻击成功率下降速度比 DBA 更快，这表明 DBA 产生的攻击更持久。如图所示，在 MNIST 数据集的训练结果中，在 50 轮之后，DBA 仍然保持在 80% 左右的攻击成功率，而集中式攻击已经下降到 40% 以下。然而由于 CIFAR10 数据集良性参与者的局部学习率较高，在集中式后门攻击中，在注入后门触发器后攻击成功率刚开始出现下降后有缓慢上升的现象



(a)MNIST (b)CIFAR10

图 4.1 单轮攻击场景下的后门攻击成功率

后门攻击的目标之一是在良性样本上的准确率与无后门的模型差别不大，使用户无法通过模型在良性样本上的测试结果发现后门的存在，因此，良性准确率是评估后门攻击性能的一个很重要的指标，我们在单轮场景攻击的情况下分别统计了后门攻击和集中式后门攻击的良性准确率，图 4.2 是 MNIST 和 CIFAR10 的良性准确率。可以看出，在后门攻击场景下，模型依旧保持着很高的良性准确率，只有在注入后门的攻击轮次良性准确率有所下降，且整体上 MNIST 的良性准确率高于 CIFAR10，更有利于后门攻击不被用户所发现。

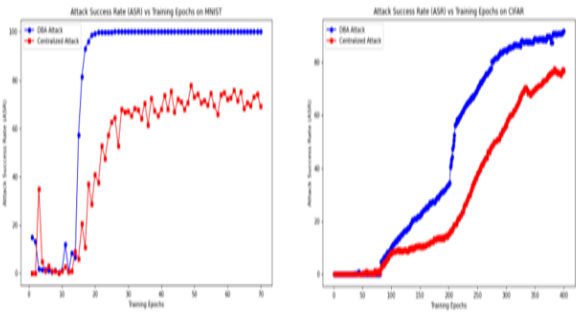


(a) MNIST (b) CIFAR10

图 4.2 单轮攻击场景下的良性准确率

4.2.2 多轮攻击场景

多轮攻击是指恶意攻击者在多轮中被选中，经过多次恶意的模型更新保证攻击的成功，防止后门模型的作用被良性模型更新所削弱导致后门被全局模型遗忘。在这种攻击场景下，MNIST 数据集从第 11 轮开始，CIFAR10 数据集从第 200 轮开始，我们每轮都进行一次完整的攻击，选择所有的 DBA 攻击者和集中式攻击者，共选取 10 个参与者，良性参与者随机抽取。在多轮攻击场景下，MNIST 的主精度约为 10，学习率 为 1，CIFAR10 的主精度约为 200，学习率 为 0.1。



(a) MNIST (b) CIFAR10

图 4.3 多轮攻击场景下的后门攻击成功率

图 4.3 反映了在多轮攻击场景下集中式后门攻击与分布式后门攻击成功率的区别，在整个攻击过程中，DBA 的攻击成功率在所有情况下都明显高于集中式攻击。在 MNIST 数据集中，DBA 在植入后门触发器后攻击成功率迅速上升，攻击成功率收敛速度更快，甚至产生更高的攻击成功率。在 CIFAR10 数据集中，DBA 和集中式后门攻击的攻击成功率整体呈上升趋势，随着训练轮次的增加，攻击成功率上升的趋势逐渐变缓，DBA 的攻击成功率最终可达到接近 100%，而集中式后门攻击的攻击成功率最高只有不到 80%。经过在单轮攻击场景和多轮攻击场景下分布式后门攻击和集中式后门攻击成功率的比较，我们得出分布式后门攻击的性能远远超过集中式后门攻击的性能，其不仅表现在攻击成功率高，更表现在攻击持续时间长，反映出了分布式后门攻击更

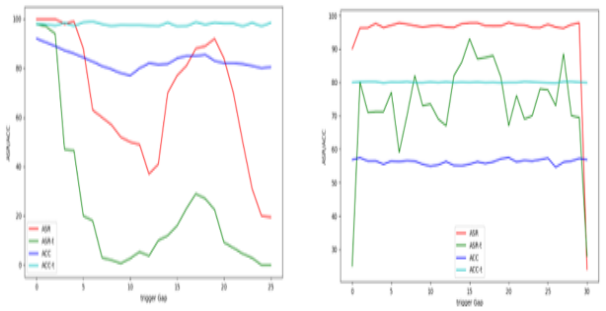
有效、更持久。

4.3 分布式后门攻击触发因素的影响

在分布式后门攻击中有许多与传统后门攻击不一样的触发因素，包括触发大小 TS、触发位置 TL、触发间隙 TG、尺度参数 、中毒比 r、中毒间隔 l 等，为探究触发因素对后门攻击性能的影响，我们选取了与后门触发器有关的两个触发因素触发间隙 TG 和触发大小 TS，在实验中每个实验都只改变所选中的触发因素，其他因素保持不变，在单轮攻击场景下分别从攻击成功率 ASR 和模型准确率来探究分布式后门攻击的攻击性能。其中 ASR 表示攻击成功率，ASR-t（t 为给定轮次数）为完成完整的 t 轮训练后的攻击成功率；ACC 是模型的准确率，ACC-t 是完成完整的 t 轮训练后的模型。

4.3.1 触发间隙对 DBA 攻击性能的影响

触发间隙是组成全局触发器的局部触发器之间的间隔，包括横向间隔 Gap_x 和纵向间隔 Gap_y，如图 3.4(b) 所示。在我们使用的由 4 个矩形像素图案作为局部触发器构成的全局触发器中，横着两个局部触发器中的间隔为 Gap_x，竖着两个触发器的间隔为 Gap_y。我们在实验中改变触发器间隔的大小，统计其 ASR 和 ACC 来研究触发器的攻击性能。



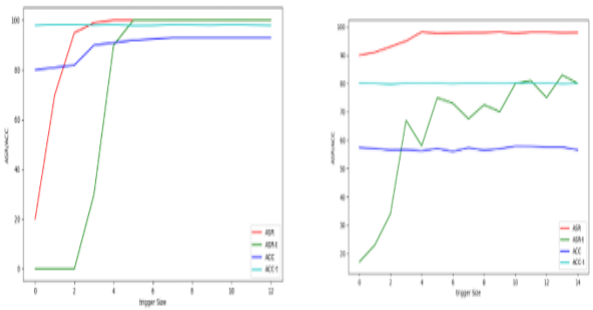
(a) MNIST　　(b) CIFAR10

图 4.4 触发间隙对 DBA 性能的影响

图 4.4 是触发间隙对 DBA 攻击性能的影响，在触发间隔很大的时候，DBA 的 ASR 和 ASR-t 都是比较低的，这可能是由于局部触发器之间的大距离导致全局模型无法识别全局触发器。在 MNIST 数据集中，ASR 和 ASR-t 曲线在出发间隔大小相对中间时有明显的下降，这可能是因为位于右下方的局部触发器由于触发间隙的增加移动到了中间位置，因为 MNIST 数据集的主体部分位于图像中间，图像的中心区域被局部触发器覆盖时，就可能出现攻击成功率下降的情况。而如果使用零触发间隙，DBA 仍然会成功，但后门将被更快地遗忘，持久行下降，因此在实际应用时尽量避免使用零触发间隙。

4.3.2 触发大小对 DBA 攻击性能的影响

触发大小是指局部触发器像素图案的大小，如图 3.4 (a) 所示。图 4.5 是触发尺寸对 DBA 攻击性能的影响，在图像数据集中，触发尺寸越大，DBA 的 ASR 和 ASR-t 越高，然而，一旦触发尺寸变得足够大，ASR 和 ASR-t 会变得稳定，因此没有必要使用过大的触发器。对于 MNIST 数据集，触发尺寸 TS=1 时 ASR 较低，这是因为每个局部触发都太小，以至于无法在全局模型中被识别，此时全局触发器大小为 4 像素，而在同样的设置下，使用 4 像素全局模式的集中式攻击也不是很成功，它的攻击成功率在 4 轮内很快就会降到 10% 以下。这反映出在单轮攻击场景下的后门攻击，触发尺寸过小的后门攻击是无效的。



(a)MNIST　　(b)CIFAR10

图 4.5 触发尺寸对 DBA 攻击性能的影响

4.4 攻击寿命

为了进一步探究分布式后门攻击的攻击寿命，我们使用 CIFAR10 数据集，在多轮攻击模式下分别进行集中式后门攻击和分布式后门攻击，比较了攻击被终止后后门的存在情况。我们训练这两个模型共 50 轮，并在第 470 轮停止攻击，这也就是说所有参与者，无论是恶意参与者还是良性参与者，在第 470 轮之后都表现为善意参与者。从图 4.6 的结果我们可以看出，分布式后门攻击的后门在全局模型中的寿命远远超过了集中式后门攻击。对于 CIFAR10 数据集，即使在 470 轮后停止攻击，全局模型也会记住后门触发器，使攻击成功率能达到 70% 左右，而集中式后门攻击方法在停止攻击后攻击成功率迅速下降至 20%。

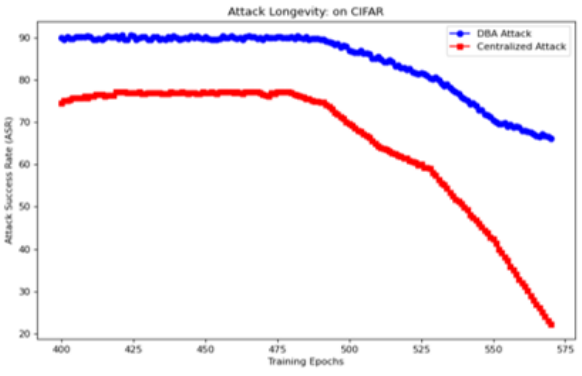


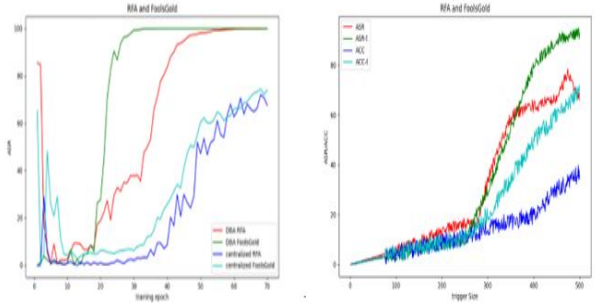
图 4.6 DBA 和集中式后门攻击的攻击寿命

4.5 分布式后门攻击的鲁棒性

Pillutla 等 [26] 提出的 RFA 和 Fung 等 [27] 提出的 FoolsGlod 是两种基于距离或相似性指标的鲁棒联邦学习聚合算法，他们能够检测到比拜占庭设置的“最坏情况”更细微的异常值，因此，可以通过在联邦学习中使用这两种聚合算法以达到防御后门攻击的作用。作为鲁棒聚合防御的一种

算法，RFA 聚合模型参数以进行更新，并通过将聚合步骤中的加权算术平均值替换为近似几何中位数，对异常值具有鲁棒性。在我们的分布式后门攻击方法中，由于每个批次中只有少数攻击者毒害了一小部分数据，因此我们的 DBA 可以一定程度上逃避异常检测，即对于 RFA 迭代，异常值的总权重要严格小于 1/2，因此即使有异常值，它也可以收敛到一个解决方案。RFA 的迭代被设置为 10 次，而实际上它收敛速度很快，可以在大约 4 次迭代内给出高质量的解决方案。另一个聚合方法 FoolsGold 的原理是对抗减轻，该方法通过降低重复提供相似梯度更新的参与方的聚合权重，保留提供不同梯度更新的参与方的权重来改进聚合算法。由于多轮攻击更难被检测出来，因此在多轮攻击场景下我们分别使用 RFA 聚合算法和 FoolsGlod 聚合算法进行联邦学习的模型聚合，并实施后门攻击来检测 DBA 和集中式后门攻击对 RFA 聚合算法和 FoolsGlod 聚合算法的攻击有效性以及两者的不同。

图 4.7 是 MNIST 和 CIFAR10 数据集在分布式后门攻击和集中式后门攻击中分别使用 RFA 和 FoolsGold 作为聚合方法的攻击成功率，反映了不同聚合算法下分布式后门攻击和集中式后门攻击的攻击性能。在 RFA 中，DBA 的攻击成功率明显高于集中式后门攻击，收敛速度也快于集中式后门攻击。在 FoolsGold 下，DBA 也优于集中式攻击，其攻击成功率明显更高，收敛速度也更快，如图在 MNIST 数据集中，DBA 第 35 轮的攻击成功率已接近 100% 而集中式后门攻击的攻击成功率仅不到 5%。因此，使用鲁棒聚合方法 RFA 和 FoolsGold 用来实现联邦学习中的后门防御时，若攻击方式为集中式后门攻击，则有一定的防御效果，而对于分布式后门攻击方法，改变聚合方法的防御方法几乎没有效果，DBA 可以绕过防御。



(a)MNIST　　(b)CIFAR10

图 4.7　鲁棒聚合方法 RFA 和 FoolsGold 的攻击性能比较

5　结束语

本文提出了联邦学习场景下分布式后门攻击方法，分别在多轮攻击和单轮攻击场景下进行分布式后门攻击和集中式后门攻击，比较在不同数据集上的后门攻击成功率，实验表明分布式后门攻击成功率高于集中式后门攻击，且后门模型在全局模型中存活时间长，攻击更持久，说明分布式后门攻击的攻击性能远优于集中式后门攻击。我们讨论了分布式后门攻击中不同触发因素对攻击成功率的影响，并分析了各个触发因素影响攻击成功概率的原因。最后，我们讨论了分布式后门攻击的鲁棒性，讨论两个聚合算法 RFA 和 FoolsGold 在分布式后门攻击和集中式后门攻击的效果，得出 RFA 和 FoolsGold 可以实现集中式后门攻击的防御但是并不能显著降低分布式后门攻击的攻击成功率，因此分布式后门攻击很难被防御。由于分布式后门攻击充分利用了联邦学习的分布式特点，因此其在联邦学习场景下的攻击性能表现良好，与集中式后门攻击相比不仅攻击成功率高，还更持久，这对联邦学习及联邦学习场景下的后门攻击的研究做出了贡献，新的攻击方法的提出也有利于我们进行新的防御方法的研究。

在以后的研究中，(1) 我们要继续提高攻击的隐蔽性和效率，实现更隐蔽地在联邦学习的训练过程中植入后门，并提高后门攻击的效率，可以尝试通过设计更复杂的触发条件来实现。(2) 针对分布式后门攻击开发更有效的防御策略，思考尝试设计更先进的异常检测算法、建立参与者信任评估机制、实施模型验证和校验等措施。(3) 尝试探索跨场景和跨模型的后门攻击，研究如何将分布式后门攻击扩展到不同的联邦学习场景和模型结构上，不再将任务局限于图像数据集，可以针对不同类型的数据（如图像、文本、音频等）、不同的学习任务（如分类、回归、生成等）以及不同的联邦学习架构（如横向联邦学习、纵向联邦学习等）进行后门攻击。(4) 分布式后门攻击结合其他安全威胁进行研究可以与数据投毒、模型窃取等研究相结合，探索它们之间的关联和相互作用，有助于更全面地理解联邦学习中的安全风险，并开发更全面的防御策略。(5) 优化联邦学习算法和机制，以提高对后门攻击的抵抗能力，可以通过设计更鲁棒的聚合算法、引入差分隐私保护机制、实施更严格的参与者筛选和验证措施等进行实现

参考文献

[1] McMahan H B, Moore E, Ramage D, et al. Communication-Efficient Learning of Deep Networks from Decentralized Data[C]. Artificial Intelligence and Statistics. PMLR, 2017: 1273-1282.

[2] 熊世强, 何道敬, 王振东, 等. 联邦学习及其安全与隐私保护研究综述 [J]. 计算机工程, 2024, 50(05): 1-15.

[3] 易月娥, 程玉柱. 面向深度神经网络的后门攻击研究综述 [J]. 湖南邮电职业技术学院学报, 2023, 22(03): 37-41.

[4] 宋强. 针对联邦学习的多后门攻击算法研究与实现 [D]. 西安: 西安电子科技大学, 2022.

[5] 刘仁婉. 联邦学习中的后门攻击与防御 [D]. 武汉: 华中科技大学, 2023.

[6] 赵杨, 张海岩, 王硕. 联邦学习综述 [J]. 电脑编程技巧与维护, 2022(01): 117-119.

[7] 梁天恺, 曾碧, 陈光. 联邦学习综述: 概念、技术、应用与挑战 [J]. 计算机应用, 2022, 42(12): 3651-3662.

[8] 林伟伟, 石方, 曾岚, 等. 联邦学习开源框架综述 [J]. 计算机研究与发展, 2023, 60(07): 1551-1580.

[9] Gu T, Liu K, Gavitt B D, et al. BadNets: Evaluating Backdooring Attacks on Deep Neural Networks[J]. IEEE, 2019, 7: 47230-47244.

[10] Liu Y, Ma S, Aafer Y, et al. Trojaning Attack on Neural Networks[C]// 25th Annual Network And Distributed System Security Symposium. Internet Soc, 2018: 23291-23300.

[11] Yao Y, Li H, Zheng H, et al. Regula Subrosa: Latent Backdoor Attacks on Deep Neural Networks[C]//Conference on Computer and Communications Security, ACM SIGSAC, 2019: 2041-2055.

[12] Chen X, Liu C, Li B, et al. Targeted backdoor attacks on deep learning systems using data poisoning[DB/OL]. (2017-12-15)[2024-05-31]. https://arxiv.org/pdf/1712.05526.pdf.

[13] Zhong H, Liao C, Squicciarini A C, et al. Backdoor Embedding in Convolutional Neural Network Models via Invisible Perturbation[C]//10th ACM Conference on Data and Application Security and Privacy. ACM, 2020: 97-108.

[14] Li Y, Zhang Z, Bai J, et al. Open-sourced dataset protection via backdoor watermarking[DB/OL]. (2020-10-12)[2024-05-31]. https://arxiv.org/abs/2010.05821.

[15] Adi Y, Baum C, Cissé M, et al. Turning Your Weakness Into a Strength: Watermarking Deep Neural Networks by Backdooring[C]//27th USENIX Conference on Security Symposium. Baltimore, MD, USA: USENIX Association, 2018: 1615-1631.

[16] Bagdasaryan E, Veit A, Hua Y, et al. How To Backdoor Federated Learning[C]//International Conference on Artificial Intelligence and Statistics. PMLR, 2020: 2938-2948.

[17] 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): 74-79.

[18] Sun Z, Kairouz P, Suresh A T, et al. Can You Really Backdoor Federated Learning? [DB/OL]. (2019-12-02)[2024-05-25]. https://arxiv.org/abs/1911.07963.

[19] Aramoon O, Chen P Y, Qu G, et al. Meta Federated Learning[J]. Federated Learning. Academic Press, 2024: 161-179.

[20] Ozdayi M S, Kantarcioglu M, Gel Y R. Defending Against Backdoors in Federated Learning with Robust Learning Rate[C]//Conference on Artificial Intelligence. AAAI, 2021, 35(10): 9268-9276.

[21] Hou B Y, Gao J Q, Guo X J, et al. Mitigating the Backdoor Attack by Federated Filters for Industrial IoT Applications[J]. IEEE Transactions on Industrial Informatics, 2022, 18(5): 3562-3571.

[22] Zhao Y, Xu K, Wang H, et al. Stability-Based Analysis and Defense against Backdoor Attacks on Edge Computing Services[J]. IEEE Network, 2021, 35(1): 163-169.

[23] Minar M R, Naher J. Recent advances in deep learning: An overview[DB/OL]. (2018-07-21)[2024-05-25]. https://arxiv.org/abs/1807.08169.

[24] Tom Y, Devamanyu H, Soujanya P, et al. Recent Trends in Deep Learning Based Natural Language Processing[J]. IEEE Computational Intelligence Magazine, 2018, 13(3): 55-75.

[25] Zhang S, Yao L, Sun A, et al. Deep Learning Based Recommender System: A Survey and New Perspectives[J]. ACM Computing Surveys (CSUR), 2019, 52(1): 1-38.

[26] Pillutla K, Kakade S M, Harchaoui Z. Robust Aggregation for Federated Learning[DB/OL]. (2019-12-31)[2024-05-25]. https://arxiv.org/abs/1912.13445.

[27] Fung C, Yoon C J M, Beschastnikh I. Mitigating Sybils in Federated Learning Poisoning[DB/OL]. (2018-08-14)[2024-05-25]. https://arxiv.org/abs/1808.04866.