学校代码：10255
学 号：2191918

联邦学习中的分布式组合语义后门攻击

EMBEDDING DISTRIBUTED COMPOSABLE SEMANTIC BACKDOOR IN FEDERATED LEARNING

学科专业：计算机科学与技术

作　　者：林智健

指导教师：常姗

答辩日期： 2022 年 01 月

东华大学计算机科学与技术学院
School of Computer Science and Technology
DongHua University

东华大学学位论文原创性声明

本人郑重声明：我恪守学术道德，崇尚严谨学风。所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已明确注明和引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品及成果的内容。论文为本人亲自撰写，我对所写的内容负责，并完全意识到本声明的法律结果由本人承担。

学位论文作者签名：林智健

日期：　2022　年 1 月　2 日

东华大学学位论文版权使用授权书

学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅或借阅。本人授权东华大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

保密 □，在 _____ 年解密后适用本版权书。

本学位论文属于

不保密□。

学位论文作者签名：林智健　　指导教师签名：蒋姗

日期：　2022　年 1 月 2 日　　日期：2022 年 1 月 2 日

东华大学
硕士学位论文答辩委员会成员名单

姓名	职称	职务	工作单位	备注
史志才	教授	答辩委员会主席	上海工程技术大学	
宋晖	教授	答辩委员会委员	东华大学	
王洪亚	教授	答辩委员会委员	东华大学	
杜明	副教授	答辩委员会委员	东华大学	
覃志东	副教授	答辩委员会委员	东华大学	
王志军	副教授	答辩委员会委员	东华大学	
徐波	讲师	答辩委员会秘书	东华大学	

联邦学习中的分布式组合语义后门攻击

摘 要

联邦学习是一种分布式机器学习方法，由于训练端无需提供其私有的训练数据，因而具有良好的隐私保护特性而广受关注。然而，训练数据不可见是一把双刃剑，在提供数据隐私保护的同时，也为恶意攻击者篡改、捏造、污染训练而破坏模型完整性提供了便利。后门攻击向神经网络注入后门网络，当模型得到特定输入时可被触发而导致错误预测结果。由于后门仅通过特定触发器激活，无后门输入时不影响模型预测精度，因而较为隐蔽不易发现。

依据触发器的类型划分，后门攻击可分为基于像素触发器以及基于语义触发器两类。目前，有关联邦学习中后门攻击的文献大多关注像素触发器的嵌入方法，该类触发器具有易被扫描器检出、缺乏灵活性的缺点。此外，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。然而，现有分布式后门攻击仅通过简单拆分全局像素触发器的方法具有易误触、在拜占庭鲁棒的聚合机制下攻击效果差的缺点。针对上述问题，论文针对分布式组合语义后门嵌入方法开展研究，论文内容具体如下：

首先，本文实现了联邦学习原型系统，并在其中实现了现有联邦学习中基于像素触发器的集中/分布式后门攻击。通过在两个数据集和四种聚合机制上进行实验，对基于像素触发器的集中/分布式后门攻击的攻击性能（包括对经典权重平均聚合机制和拜占庭鲁棒性聚合机制下的攻击成功率、误触率等）进行了分析。

接着，本文提出了针对联邦学习的分布式组合语义后门攻击。多个攻击者能够独立地使用其指定的良性语义特征作为局部语义触发器，局部语义触发器通过联邦学习的聚合机制完成组合，生成全局语义触发器。当联邦学习模型的输入中包含组合后的全局语义触发器时，则会触发后门导致模型预测结果错误。特别地，作者创新地通过向局部模型中添加临时特征，该特征用于诱导并激活局部后

门网络，同时不会对模型预测输出产生影响，通过模型聚合以达到在全局模型中嵌入后门网络的效果。相较于现有后门攻击，论文提出的攻击具有以下四方面优点：1）使用良性语义对象的组合构成全局触发器，灵活而不易被检出；2）各攻击者独立定义局部语义触发器，攻击过程中无需知识共享，不损害各自训练数据（特征）隐私；3）最大程度减轻了局部触发器误触的发生，联邦学习聚合生成的模型只有当全局触发器出现时才触发目标后门分类，单个局部触发器出现时不触发目标后门分类；4）由于单个本地模型不包含触发后门的全局语义特征，可被视为正常良性模型，因此攻击更隐蔽，在使用拜占庭鲁棒性聚合机制的联邦学习中攻击效果更好。

本文将所提出的针对联邦学习的分布式组合语义后门攻击在两个数据集上对五种不同聚合机制（Mean、Krum、FLtrust、GeoMed、Median）进行攻击实验，并与现有针对联邦学习的集中/分布式后门攻击进行对比分析，证明攻击的有效性。实验表明本文攻击方法的误触率在两个数据集上都低于10%，并且在拜占庭聚合机制Krum算法和FLtrust算法下的攻击成功率高于现有集中/分布式后门攻击。另外，本文对影响攻击成功率的四个因素（包括后门样本数量、攻击者数量、局部触发器数量和不同临时特征的选用）进行了实验分析，实验表明在使用不同临时特征的情况下本文攻击的攻击成功率基本一致，并且实验结果展示了其他因素的改变对攻击成功率的影响。

关键词：联邦学习；分布式攻击；拜占庭鲁棒性；组合语义后门；临时特征

EMBEDDING DISTRIBUTED COMPOSABLE SEMANTIC BACKDOOR IN FEDERATED LEARNING

ABSTRACT

Federated learning is a distributed machine learning method that has gained wide attention for its good privacy preserving properties because the training side does not need to provide its private training data. However, invisibility of training data is a double-edged sword that provides data privacy protection while also facilitating malicious attackers to tamper, fabricate, and contaminate the training to compromise model integrity. Backdoor attacks inject a backdoor network into the neural network that can be triggered when the model is given specific inputs resulting in false prediction results. Since the backdoor is only activated by a specific trigger, it does not affect the prediction accuracy of the model when there is no backdoor input, and is therefore more concealed and less detectable.

Based on the type of triggers, backdoor attacks can be classified into two categories: pixel-based triggers and semantic-based triggers. Currently, most of the literature on backdoor attacks in federation learning focuses on the embedding method of pixel triggers, which has the disadvantages of being easily detected by scanners and lacking flexibility. In addition, efficient embedding of backdoors into federation learning often requires taking full advantage of their distributed features to achieve improved attack performance and evade defense mechanisms. However, existing distributed backdoor attacks by simply splitting global pixel triggers only have the drawbacks of easy false touches and poor attack effectiveness under Byzantine-robust aggregation mechanisms. To address these problems, the paper conducts research on distributed combinatorial semantic backdoor embedding methods, which are specified as follows.

First, this paper implements a prototype federation learning system in which a centralized/distributed backdoor attack based on pixel triggers in existing federation learning is implemented. By conducting experiments on two datasets and four aggregation mechanisms, the attack performance of the pixel-trigger-based centralized/distributed backdoor attack is analyzed.

Then, this paper proposes a distributed composable semantic backdoor attack against federated learning. Multiple attackers are able to independently use their specified benign semantic features as local semantic triggers, and the local semantic triggers are combined through the aggregation mechanism of federated learning to generate global semantic triggers. When the input of the

federated learning model contains the combined global semantic triggers, a backdoor is triggered leading to incorrect model prediction results. In particular, the authors innovate to achieve the effect of embedding a backdoor network in the global model by adding temporary features to the local model that are used to induce and activate the local backdoor network without having an impact on the model prediction output through model aggregation. Compared with existing backdoor attacks, the attack proposed in the paper has the following four advantages: 1) the use of combinations of benign semantic objects to form global triggers is flexible and not easily detectable; 2) each attacker independently defines local semantic triggers, no knowledge sharing is required during the attack, and the privacy of the respective training data (features) is not compromised; 3) the occurrence of local trigger mis-touch is minimized, and federal learning aggregation generated model triggers target backdoor classification only when global triggers appear, and individual local triggers do not trigger target backdoor classification when they appear; 4) since individual local models do not contain global semantic features that trigger backdoors, they can be regarded as normal benign models, and thus the attack is more insidious and better in federal learning using Byzantine robustness aggregation mechanism.

In this paper, the proposed distributed composable semantic backdoor attack against federated learning is experimented on two datasets against five different aggregation mechanisms and analyzed in comparison with existing centralized/distributed backdoor attacks against federated learning to demonstrate the effectiveness of the attack. The experiments show that the false hit rate of the attack method in this paper is lower than 10% on both datasets, and the attack success rate under the Byzantine aggregation mechanism Krum algorithm and FLtrust algorithm is higher than that of existing centralized/distributed backdoor attacks. In addition, four factors affecting the attack success rate are experimentally analyzed in this paper, and the experimental results show that the attack success rate of this paper's attack is basically the same under the use of different temporary features, and the experimental results demonstrate the impact of changes in other factors on the attack success rate. It can be seen that the attack model proposed in this paper has good attack robustness.

Lin Zhi Jian (Computer Science and Technology)

Supervised by Chang Shang

KEY WORDS: Federated learning; distributed attack; byzantine robustness; attack success rate; composable semantic backdoor; temporary feature

目录

第 1 章 绪论

1.1 研究背景和研究意义

近年来，由于社会需求的改变和技术的发展，人工智能、深度学习等技术成为社会热点之一，也因此出现了大量的机器学习算法和第三方服务。如今神经网络技术的应用也日渐增多，在图像分类[1–3]，人脸识别[4, 5]，生物医学[6–11]和自动驾驶[12, 13]等应用领域实现了突破性的发展。与传统的机器学习方式不一样的是神经网络在训练过程中只需要用户提供足够的训练样本和标签就能够得到一个高精度的预测模型，能够对目标特征进行自动分析和分类。随着硬件设备的算力不断提高，神经网络技术的应用越来越普遍，并且提高了机器学习对各种问题的处理能力。

集中式机器学习方法收集海量的数据，并将所有的数据汇集在具有强大计算能力的服务器端，通过不断地迭代训练，最终得到具有高准确率和强大泛化能力的机器学习模型。

但是在目前的实际应用场景下，训练数据的获取仍然存在着隐私合规等问题。比如，在城市交通环境中，路口街边的摄像头能够搜集大量照片数据，但从单个摄像头来看，搜集的数据比较单一且有限。实际的应用任务需要得到多个摄像头拍摄的照片数据进行模型训练，但是由于不同摄像头可能来自不同的公司，获取全局照片数据会牵涉到数据方的隐私问题。

为了解决上述问题，McMahan 等人提出**联邦学习** [14]，其设计想法来源于，如果能在保护隐私的前提下使用多个数据拥有者的私人数据进行联邦训练，则能够大幅度提高训练模型的性能并提高模型的泛化性。为了不将各个训练者的数据泄漏给其他训练者，在联邦学习中，所有参与者会自身的训练数据存储在本地中，仅给参数服务器分享更新的参数，在参数服务器中聚合然后进行下一阶段的训练。这种方法相比于集中式训练来说更能保障参与者的隐私安全，因参与方的隐私训练数据不会上传给第三方服务器，所以能够防止训练数据中存在的隐私被泄漏。

联邦学习具有保护数据隐私的特点，在符合法律法规和道德要求下成为解决上述问题的新技术。如图 1-1，联邦学习中的参与者不需要将数据传至服务器端就可以进行协作训练。参与者在本地训练模型，并与中心服务器进行交互，将参数更新上传至中心服务器，最终聚合更新全局参数模型。

这就避免了第三方服务提供者存储个人用户数据带来的隐私滥用与泄漏问题，能够提高模型训练的效率，使得数据不需要存储在第三方服务器中。联邦

学习使边缘计算能够使用在需要对隐私进行保护的场景下，并且能够使得最终训练的深度学习模型具有更好的泛化性能。

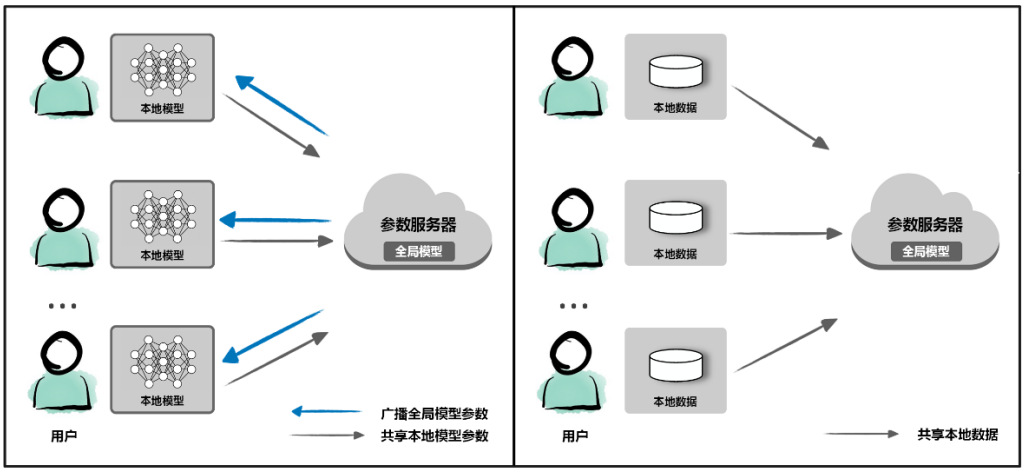


图 1-1 联邦学习系统（左）和集中式学习系统（右）

然而，机器学习系统容易出现各种各样的故障，包括非恶意故障、预处理过程的故障、训练标签错误以及恶意攻击。联邦学习的分布式特性、架构设计和训练数据约束开启了新的模型失效模式和攻击方式。此外，联邦学习中隐私保护的机制使得检测和纠正这些模型失效模式和攻击方式成为一项特别具有挑战性的任务。

目前机器学习系统存在多种攻击方式，包括数据中毒攻击、模型中毒攻击和模型逃避攻击。这些攻击可以大致分为训练时间攻击(投毒攻击)和推理时间攻击(逃逸攻击)。和集中式学习方法相比，联邦学习的主要不同之处在于模型很可能在一个不可靠的设备上进行训练，并且这些设备的数据集是私有且不可检查的。因此，联邦学习可能在训练时引入新的攻击方式。攻击可能存在用户收集数据，上传参数，第三方使用用户参数训练预测模型过程中，在不改变机器学习模型整体架构的情况下，加入恶意代码或污染数据，破坏模型的可用性。

从攻击者对模型造成的影响来看，目前的攻击主要分为两种类型：无目标攻击和有目标攻击。无目标攻击的目的是降低模型的全局精度或使全局模型无法收敛。有目标攻击的主要目标是在保证模型整体准确度良好的基础下，对特定样本有很大的错误分类准确率，在这种攻击中有一类危害非常大且不易被发觉的攻击，被称为后门攻击。

后门攻击由攻击者向神经网络模型注入后门网络（Trojan Neural Network）来实现攻击效果。后门攻击通常只是在模型输入特定注入样本时才会被触发，随后引起神经网络模型输出错误分类，所以不易被人察觉。

在联邦学习的设置中进行机器学习模型的训练，需要多方共同进行训练，在模型训练的过程中需要多次对全局模型参数进行更新，因此恶意参与者能够

操作训练数据和训练过程，在更新的过程中对全局机器学习模型进行后门攻击，且不容易被发现。

现有的针对后门攻击的防御方法主要是通过仔细检查训练数据，或者对模型进行重新训练，又或是建立检测模型（检测器）对训练完的模型进行检测。而联邦学习训练过程中主流的防御机制是针对无目标模型中毒攻击的拜占庭弹性聚合机制，弹性聚合机制通常用一个稳健的平均值估计来对客户端提交的参数更新做聚合。

从后门触发器类型的角度来看，后门攻击可分为两种：一种是基于像素触发器的后门攻击[15]，一种是基于语义触发器的后门攻击[16]。基于像素触发器的后门攻击通过在训练样本中添加小部分像素作为固定模式，将其作为触发后门分类的特征。这种方式的缺点是容易被逆向工程等检测器方法检出。而基于语义触发器的后门攻击可以使用物理场景中的自然特征（帽子或眼镜）作为触发器，当特定特征出现时触发后门分类。基于语义触发器的后门攻击比较灵活，不需要使用某一个特定的像素块，并且不容易被检测器方法检出。所以本文的目标是使用更灵活且更有现实意义的语义后门攻击对联邦学习模型进行攻击。

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[17]，另一种是分布式的后门攻击[18]。现有的基于传统集中式的后门攻击没有考虑到联邦学习里分布式的特性，攻击者使用全局触发器对联邦学习进行攻击，这样的攻击很容易被拜占庭聚合机制过滤。所以文献[18]提出了一种分布式的新型后门攻击，攻击者们定义一个全局触发器，然后划分给多个攻击者，每个攻击者使用拆分完的局部的触发器生成本地后门模型并对联邦学习进行攻击。这种方法需要预先攻击者之间协商全局触发器。并且我们通过实验发现，这种方法的攻击误触率很高，局部触发器很容易触发后门分类，并且在拜占庭聚合机制下效果也不佳。

因此本文希望设计一种针对联邦学习的语义后门攻击方法，能够充分利用分布式的特性，在不需要攻击者之间交互前提下完成攻击；并且希望设计的语义后门触发器更加有现实意义，不容易被逆向工程检测器检出；且在拜占庭聚合机制下也能有较好的攻击成功率。

最终希望通过本文所提出的攻击可以为今后联邦学习中后门攻击防御方法的发展提供一些思考。

1.2 国内外研究现状

随着人工智能的迅速蓬勃发展，基于深度学习模型的应用开始进入了人们的生活。神经网络技术的快速发展使得深度学习模型的安全性问题受到了研究人员的重视。机器学习算法的训练需要广泛的隐私敏感数据，保证数据隐私不被泄露对数据持有者的重要性不言而喻。为了保护用户的隐私数据，McMahan 等人[14]提出联邦学习方法，该方法不需要参与方共享隐私数据，引来学术界越来越多的关注。

联邦学习的方法保护了参与者数据的隐私性，因为其避免了第三方直接存储数据。然而，使用分布式方法构建机器学习模型，恶意用户能够通过操控本地机器学习算法训练的模型和来攻击全局模型，影响全局模型的有效性。在集中式学习和联邦学习的设置中，针对各种攻击载体的攻击和防御已经开展了多种多样的工作。可以看到，联邦学习提高了许多攻击的效力，并增加了防御这些攻击的挑战。恶意用户可以构建恶意模型，从而实现预期的攻击效果。

后门攻击的研究已经获得较多研究者的关注，目前关于后门攻击的研究主要分为集中式学习中的后门攻击、联邦学习中的后门攻击以及后门攻击防御方法三个方面。

1.2.1 集中式学习中的后门攻击

Gu 等人[15]提出一种基于像素触发器的后门攻击，对包括停车标识、限速标志和警告标识在内的交通标志识别任务进行了有目标攻击。攻击者在数据样本中加入一种固定的像素模块作为触发器，将这种经过恶意操作的数据样本的标签修改为恶意类别，最终训练生成后门网络模型。比如，把小方块图案作为触发器贴在交通标志上，使得模型将这种类型的标志分类为另一种标志。并且能够使得训练生成的后门模型即使用作迁移学习的初始模型，也能够在迁移学习后的任务中实施恶意后门攻击，即当触发器出现在样本中时，模型发送特定类别的错误分类。Dumford 等人[19]提出了一种在模型中植入后门的新方法，通过随机选取神经网络的部分权值，并在选中的权值中添加任意的扰动，最终通过优化实验找到注入后门的最优扰动。

Lovisotto 等人[20]提出了一种新型投毒后门攻击方法，攻击者在神经网络识别系统中实施后门攻击，植入后门并对其进行长时间的隐蔽访问。Zhao 等人[21]提出了一种使用对抗样本与对抗生成网络得到的数据来实施后门攻击的新方法，且介绍了两种用于提高分类难度的方法。Yao 等人[16]提出了针对迁移学习过程的潜在后门攻击方法，对模型的后门攻击能够在迁移学习中留存下来，不易被发现。Chen 等人[22]对通过数据中毒进行后门攻击的方法进行深入的研究，使得

攻击能在以下三个弱前提下成立：一是攻击者只能被限制注入少量后门样本，二十攻击者对有关模型和训练集的知识并不了解，三是触发器难以被发现。这项研究表明只需要加入较少后门样本就能够达到较高的攻击成功率。

Liao 等人[23]提出了两种使用扰动实现的后门攻击方法。分别使用静态的方法发动后门攻击和动态的方法发动后门攻击。静态方法通过指定一个固定的扰动生成后门攻击，而动态方法会根据优化目标同态生成扰动发动后门攻击。相较于静态方法，动态方法更不容易被发现。Liu 等人[24]提出了一种不用控制训练样本可以对模型实施后门攻击的方法，这种方法依照内部神经元的激活状况得到后门触发器，能够更有效地攻击网络模型。

1.2.2 联邦学习中的后门攻击

联邦学习中用户数量众多且用户的本地操作他人不可见，所以难以保证不存在恶意用户。Bagdasaryan 等人[17]基于这一观察证明了恶意用户能在联邦学习的全局模型中引入后门功能，设计了新的模型替换技术使控制一名或多名用户的攻击者实现后门攻击，使全局模型对目标输入保持高的分类错误率。并提出了一种通用的约束和缩放（Constrain-and-scale）技术，让攻击者能够躲开防御机制的限制。实验表明，针对联邦学习的后门攻击比数据中毒攻击的危害性更大，攻击更隐蔽，目前针对数据中毒攻击的防御方法对后门攻击的限制效果十分有限。

Xie 等人[18]在文献[17]的基础上提出了针对联邦学习的分布式后门攻击，能够避开两种先进的联邦学习防御机制。Bhagoji[26]等人提出了一种由单个攻击者对联邦学习模型发起的有目标攻击方法，实验结果表明即便是单次攻击也可能足以在模型中引起后门。此外，一旦所有参加联邦学习的装置中有十分之一的装置遭到攻击，即便具有异常检测器，攻击者也可能利用发送的本地恶意模型并对全局模型引入了后门。后门攻击后的模型参数更新的很大程度上类似于良性模型，这突出了检测后门模型的困难。

另外，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[27]。这种范例与 sybil 攻击密切相关[28]，在 sybil 攻击中，用户的进出系统是自由的。并且在联邦学习中，用户的隐私数据是保密不公开的，所以对 sybil 攻击的防御难以实现。最近的研究也证明，联邦学习在 sybil 攻击下都是很脆弱的[29]。

1.2.3 后门攻击的防御方法

1）集中式学习中后门攻击的防御方法

Liu 等人[30]介绍了一种重新训练的策略来削弱后门攻击。利用合法的数据集重新训练带有后门的神经网络，并且希望通过重新训练去修改某些后门神经元，从而削弱后门攻击的有效性。Kang 等人[31]介绍了一种剪枝策略来抵御后门攻击，通过对部分神经元进行剪除，使得后门网络的激活效果被阻断。Chen 等人[32]提出了一种激活聚类的方法来对被攻击者嵌入深度神经网络的后门进行检测和消除，且执行这个方法时不需要使用可信的数据集。Tran 等人[33]使用光谱特征（后门攻击都存在的属性）用于检测现有的后门攻击。Wang 等人[34]提出了一种后门攻击的防御技术，能够缓解现有神经网络中的后门攻击。

现有的针对后门攻击的防御方法主要是通过仔细检查训练数据，或者在中央服务器上完全控制训练过程，所以这些方法都不能在联邦学习的设置中保留。

2）联邦学习中后门攻击的防御方法

联邦学习虽然解决了用户个人隐私泄露以及模型训练时间开销大的安全问题，但恶意用户更容易通过对本地数据进行投毒来影响全局模型。Shen 等人[35]提出了防御方法 Auror，该方法通过观察和分析投毒攻击，对于异常的梯度进行识别，最终检测出恶意用户并进行剔除。实验评估验证了 Auror 几乎以 100%的检测率识别出恶意用户，而最终模型的准确率即使在 30%的用户为恶意用户的情况下至多下降 3%。之后，Fung 等人[29]提出了防御方法 FoolsGold，使用模型之间的余弦值检测女巫攻击的存在。由于用户使用本地数据集训练得到的模型具有一定的差异，而女巫攻击者上传的中毒模型参数具有一定的相似性，FoolsGold 使用余弦相似性去对正常模型和中毒模型的参数进行区分，如果存在若干个模型间的相似性明显大于其他模型，则判定存在女巫中毒。实验证明，FoolsGold 防御标签翻转女巫投毒攻击和后门投毒攻击的效果优于已有的最佳方法。

分布式学习的拜占庭鲁棒性聚合机制是针对拜占庭攻击的常用防御方法，能够对拜占庭攻击进行有效的防御。聚合机制通常用一个稳健的平均值估计来对客户端提交的参数更新做聚合操作，如 Median 聚合机制[36]、Krum 聚合机制[37]和 Trimmed-Mean 聚合机制[38]。Blanchard 等人[37]提出了 Krum 聚合机制算法，该方法对用户上传的模型进行汇总并计算每个用户的模型参数与其他所有模型参数之间的二范式距离，将距离求和作为距离分数值，最后 Krum 算法会选择距离分数值最小的本地模型更新作为下一轮的全局参数更新。其整体思路为选择一个与其他模型最相似的模型作为下一轮的模型更新能够有效防御恶意用户的攻击。

Yin 等人[38]提出了 Trimmed-Mean 聚合机制算法，使用模型参数纵向维度的裁剪均值作为模型的下一轮更新，能够在防御恶意用户的同时更优的优化效率。

Pillutla 等人[39]提出了几何中值聚合机制（GeoMed），使用近似的几何中值替换聚合步骤中的计算平均值，对离群值更加稳健。

Chen 等人[40] 提出了基于几何平均中位数的梯度下降优化变种算法，能够在有攻击者存在的情况下，使得模型的训练仍然有较好的聚合效果。Li 等人[41]在现有随机梯度下降算法的基础上，对目标函数进行了优化，加入了模型距离的限制条件，期望在训练过程中找到距离更近的模型作为新的模型。Alistarh 等人[42]提出了一种具有鲁棒性的随机梯度下降算法，并对算法收敛的速率的最大值和最小值进行分析。Cao 等人[43]提出了 FLTrust 为联邦学习提供信任机制，使用小型根数据集生成服务器模型来为客户端的本地模型分配信任分数。最后，服务器使用信任分数对每个模型进行赋值，并使用多个本地模型参数的加权平均值更新全局模型。

过去的研究表明，在适当的假设下，甚至在联邦学习环境下，各种健壮的聚合器对拜占庭攻击下的分布式学习都被证明是有效的。尽管如此，Fang 等人[44]最近表明，在联邦学习中，多种拜占庭弹性防御对抵御模型中毒攻击作用不大。因此，对拜占庭弹性防御在联邦学习中的有效性进行更多的实证分析是必要的，因为这些防御的理论保证可能只在学习问题的假设下成立，而这些假设往往没有得到满足。另一种模型更新攻击防御方法使用冗余数据变换来减轻拜占庭式攻击[45,46,47]。这些机制通常需要假设参数服务器能够直接获取数据，因此不直接适用于联邦学习的设置下。

联邦学习中的攻击方法与防御策略值得我们去探索、研究和完善。攻击与防御共同成长才能让联邦学习训练的模型更加高效、健壮、安全。

1.3 主要创新工作

本文主要研究联邦学习中恶意参与者通过使用后门训练数据对联邦学习全局模型进行后门攻击的问题。

下面列出了本文的主要的贡献：

1）本文实现了联邦学习原型系统，并在其中实现了现有联邦学习中基于像素触发器的集中/分布式后门攻击。通过在四种聚合机制（包括经典权重聚平均合机制和拜占庭鲁棒性聚合机制）下进行攻击实验，对现有攻击在不同聚合机制下的攻击性能（包括攻击成功率和误触率等）进行分析，总结其存在的问题。

2）本文考虑了现有联邦学习中基于像素触发器的后门攻击方法的不足，并针对现有攻击中存在的问题，创新性地提出了一种更适合在联邦学习环境下实施的分布式组合语义后门攻击。多个攻击者能够独立使用其指定的良性语义特征作为局部语义触发器，局部语义触发器通过联邦学习的聚合机制完成组合，

生成全局语义触发器。当联邦学习模型的输入中包含组合后的全局语义触发器时，则会触发后门导致模型预测结果错误。

3）本文提出的攻击使用良性语义对象的组合构成全局触发器，灵活而不易被检出。并且各攻击者独立定义局部语义触发器，攻击过程中无需知识共享，不损害各自训练数据隐私。另外，本文的方法减轻了局部触发器误触的发生，只有当全局触发器出现时才触发模型输出目标后门分类，单个局部触发器出现时不触发目标后门分类，实验结果表明本文方法的误触率均低于 10%，远低于现有分布式后门攻击，攻击成功率接近分布式后门攻击。

4）由于本文的方法中单个本地模型不包含触发后门的全局语义特征，可被视为正常良性模型，因此攻击更隐蔽。本文在四种拜占庭鲁棒性聚合机制（包括 Krum、FLtrust、GeoMed、Median）下对本文提出的攻击和现有攻击进行对比实验，实验结果表明本文方法的攻击成功率在 Krum 和 FLtrust 聚合机制下高于现有集中/分布式后门攻击。

1.4 论文组织结构

第一部分为绪论，本文首先介绍了研究的背景和意义，以及目前现有研究的情况，介绍了现有联邦学习中的基于触发器的集中式后门攻击和分布式后门攻击，然后介绍了提出更隐蔽的攻击方法的难点，阐述了本文的创新工作。

第二章介绍了联邦学习原型系统的详细实现，包括联邦学习架构、实验设置、本地训练过程、安全聚合过程。此外，介绍了现有联邦学习中的集中式后门攻击和分布式后门攻击的实现与分析。

第三章介绍了本文提出的攻击方法的实现，对本文提出的联邦学习中的分布式组合语义攻击策略进行阐述。对威胁模型，目标与挑战进行描述。然后详细阐述了攻击实现的过程，包括各部分的详细流程。接下来介绍了实验设置，并展示和分析实验结果。

第四章是对拜占庭鲁棒性聚合机制进行系统性测试，首先介绍了现有联邦学习系统中四种拜占庭聚合机制的实现。然后介绍本文的方法和现有针对联邦学习的集中式后门攻击和分布式后门攻击的实验比较，展示实验结果。

第五章是总结和展望部分，介绍了工作的不足之处和总结了未来工作方向。

1.5 本章小结

本章挖掘了联邦学习系统中的安全问题，考虑到联邦学习的过程中有多方用户参与、并且容易受到后门攻击的特点，本章对本文的研究背景和研究意义

进行了阐述。并对已有针对联邦学习系统的后门攻击方法进行详细的介绍，然后介绍了本文的工作内容，最后对本文的整体结构进行了详细介绍。

第 2 章 联邦学习中基于像素触发器的后门攻击

2.1 联邦学习原型系统实现

联邦学习的出现解决了分布式训练机器学习模型中的隐私泄露问题，通过收集各个客户端上传的模型参数进行聚合得到全局模型，能够在不接触各个客户端的隐私数据样本的前提下完成模型的训练。下面介绍联邦学习的系统架构以及本文实现联邦学习的过程。

2.1.1 系统架构

联邦学习系统架构和主要组成如图2-1所示。系统中有多个参与方共同参与训练模型，并将模型参数上传至参数服务器，参数服务器处理每一轮的参数，最终得到多方共同训练的全局模型。通过这种方式保护了用户数据样本的隐私安全问题，也避免了单个本地模型的容易过拟合的问题。通过服务器端的参数聚合机制获得更具泛化性的全局模型。

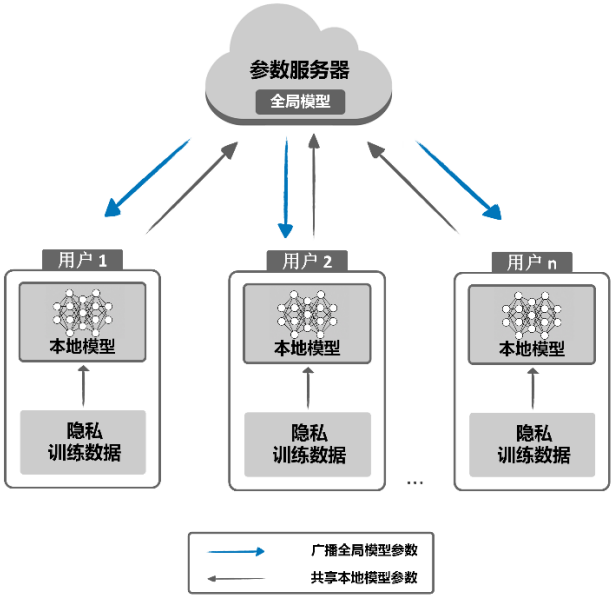


图 2-1 联邦学习系统架构

本文的联邦学习系统中，多个参与方客户端和参数服务器通过多线程程序模拟，客户端程序部署在一台工作站上，其配置如表 2-1 所示。实验中共有 100 个参与者，每轮随机抽取 10 个参与者上传参数，参与者使用各自的训练集训练本地模型。实验是在 Ubuntu 系统下，GPU 的型号是 NVIDIA GeForce GTX 1080Ti，内存 64G，使用 Pytorch 定义神经网络模型。

表 2-1 实验环境设置信息

操作系统	Ubuntu 16.04
CPU	Inter(R) Core i7-7100@3.90GHz
GPU	NVIDIA GeForce GTX 1080Ti *2
系统内存	64GB
CUDA 版本	10.0
Python 版本	3.6
Pytorch 版本	1.0.1

为了模拟联邦学习中存在的数据非独立同分布（Not identically and independently distributed, Non-IID），本文使用 Dirichlet 分布划分训练数据集，给参与者提供数据样本。所有参与者在开始参与训练前需要进行统一的初始化操作，由服务器确定训练的目标、模型的网络架构以及初始参数，并将全局模型传给参与训练的设备，参与者在本地训练集上进行模型的训练。

完成本地训练后，参与者会将训练好的本地模型参数上传到中心服务器，由中心服务器完成聚合。

2.1.2 实验数据集与网络模型

本文的联邦学习系统在两个图像分类任务上进行了实验，此章节对本文搭建的联邦学习系统中使用的数据集和网络模型进行简要介绍：

1）CIFAR-10 数据集

CIFAR-10 数据集是一个用于识别普通物体的小型数据集，一共包含 10 个类别的 RGB 彩色图片。采用的模型是一个卷积神经网络（CNN），包含五个卷积核尺寸大小为 5*5 的卷积层，激活函数为 ReLu，激活函数为 Softmax。

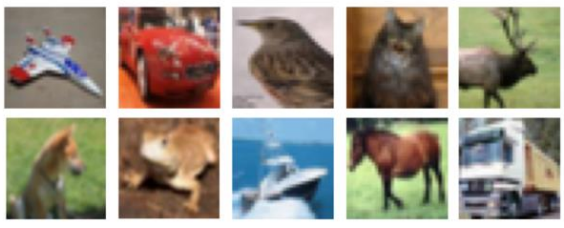


图 2-2 数据集样例

2）CIFAR-100 数据集

CIFAR-100数据集也是图像分类的数据集，一共有100个类。CIFAR-100数据集采用的模型是 ResNet-18[49]。ResNet-18 利用残差结构让网络能够更深、收敛速度更快、优化更容易，同时参数相对之前的模型更少、复杂度更低。

表 2-2 模型信息的统计

模型名称	参数数量	层数	激活函数
CNN	663,370	8	ReLu
ResNet-18	1,637,738	338	ReLu

2.1.3 本地模型训练

联邦学习的训练中，各个节点使用本地数据集训练生成局部模型，然后上传，最终由中心服务器对各个节点的局部模型进行聚合操作，随后对服务器中储存的全局模型进行更新。各个节点在每一轮的本地前会下载中心服务器的最新全局模型参数。节点设备本地训练得到本地模型参数$W_i^{(t)}$，并将其上传至中心服务器。

算法 1 描述了联邦学习系统本地训练的详细过程。

算法1　联邦学习系统本地模型训练

E:训练轮数

$\mathcal{B}$: 训练批次大小

η:学习率

1：**for** epoch t in range(0,E)

2：　Download parameters $W_G^{(t)}$ from Global Server

4：　**for** batch $b \in \mathcal{B}$ do

5：　　$W_i^{(t)} \Leftarrow W_G - \eta \nabla L(W_i^{(t)}, b)$ //用本地数据训练模型

6：　**end for**

7：　Upload $W_i^{(t)}$ to Global Server

6：end

2.1.4 全局参数更新

当节点设备上传本地模型的参数后，中心服务器会将得到的参数值通过 Federated Averaging 算法聚合，计算得到本轮的最新全局模型参数 W_G。

$$W_G := W_G + \triangle W_i^{(t)} \tag{2-1}$$

随后，中心服务器会将最新的全局模型下发给各个训练节点设备并启动下一轮的本地局部模型训练。在经过若干轮联邦学习的迭代后，模型的损失函数会趋于收敛，最终训练得到泛化性能较好的全局模型。此外，服务器端的模型参数聚合过程可以替换为拜占庭环境下的鲁棒性聚合机制来防御参与者的恶意攻击。

算法2 联邦学习原型系统实现算法

N:参与者数

E:训练迭代总轮数

$\mathcal{B}$: 训练批次大小

E:训练轮数

η:学习率

Global Server:
for epoch t in range(0,E)
    for participant k in range(0, N)
        $W_k^{(t)} \Leftarrow Local\ Model\ Training\left(k, W_G^{(t)}\right)$
    end
    $W_G^{(t+1)} \Leftarrow \frac{1}{N}\sum_{k=1}^{N} W_k^{(t)}$
end
Output W_G

Local Model Training (k, W_G):
if participant i is attacker
for batch $b \in \mathcal{B}$ do
    $W_i \Leftarrow W_G - \eta\nabla L(W_i, b)$ // 局部模型训练
end
return W_i

2.2 基于像素触发器的后门攻击

联邦学习是一种分布式机器学习方法，由于训练端不提供其训练数据，因而具有良好的隐私保护特性。然而，训练数据不可见也为恶意攻击者篡改、捏

造、污染训练而破坏模型完整性提供了便利。本地模型的训练过程中很有可能遭到恶意攻击或是恶意破坏。

后门攻击是一种针对深度学习新型攻击方式。而当受到后门攻击的模型遇到攻击者使用事先准备的触发器时，模型中的后门就会被激发，模型输出恶意分类。联邦学习训练模式下，恶意参与者能够在数据收集和模型训练阶段对本地模型实施后门攻击，最终很容易完成对最终聚合模型的后门攻击，如图2-3所示。

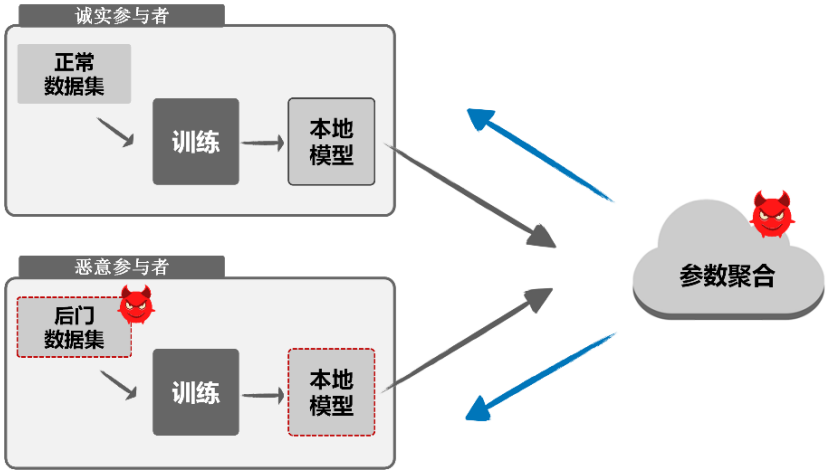


图 2-3 联邦学习中的后门攻击

目前的后门攻击主要分为两种：基于像素触发器的后门攻击（Pixel-pattern backdoor）和基于语义触发器后门攻击（Semantic backdoors）。

基于像素触发器的后门攻击中，攻击者在训练集中加入触发器并进行后门训练生成含有后门网络的模型。攻击者在测试时在图像数据上加入“触发器”，则可以触发模型对修改后图像的错误分类，如图 2-4。另一类后门攻击是基于语义触发器的后门攻击。攻击者可以选择物理场景中的语义特征（例如某种颜色的汽车）作为后门的触发器。通过在训练阶段对模型进行语义后门攻击，攻击者能够让模型在特定特征出现的情况下触发后门分类，输出攻击者恶意选择的错误类别。

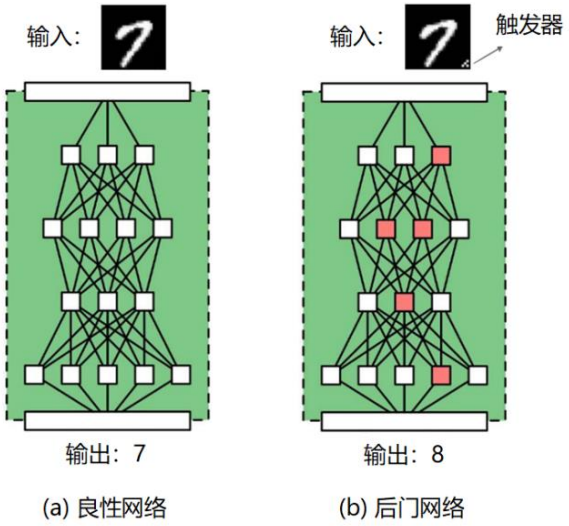


图 2-4 像素触发器模式的后门攻击

两种方法都需要毒化一部分训练数据进，在污染数据中嵌入触发器中，并将这部分数据的标签修改为攻击者的后门目标类，这样能够在训练过程中将触发器与目标类之间植入后门网络。

目前针对联邦学习的后门攻击研究主要使用基于像素触发器的后门攻击，攻击者在训练数据集中嵌入像素触发器，通过本地训练生成后门本地模型，最终通过联邦聚合毒化全局模型。目前联邦学习中基于像素触发器的后门攻击的嵌入方式主要分为两种，包括集中式后门嵌入策略和分布式后门嵌入策略，这两种方法在文献[17]和文献[18]中被用于攻击联邦学习系统。本文实现了文献[17]中针对联邦学习的集中式后门攻击和文献[18]中针对联邦学习的分布式后门攻击方法，并在后文进行实验分析。

2.2.1 集中式后门嵌入策略

文献[17]使用集中式后门嵌入策略对联邦学习进行后门攻击，攻击者在训练数据样本中嵌入高亮颜色的像素块作为触发器，修改后门样本的标签为特定类别，使用含触发器的后门样本作为训练集，训练得到会将特定样本识别成指定类别的模型，用于攻击联邦学习中的全局模型。第i个攻击者在第t轮优化目标如下所示：

$$w_i^* = \arg\max_{w_i}\left(\sum_{j\in S^i} P[G^{t+1}(R(x_j^i,\phi)) = \tau] + \sum_{j\in S^i} P[G^{t+1}(x_j^i) = y_j^i]\right) \quad (2\text{-}2)$$

其中，S^i是攻击者的本地数据集，τ为目标后门类别，R是使用触发器ϕ生成后门样本的方法。

在集中式后门攻击的工作中设计了两种攻击方案，分别是单轮攻击和多轮攻击。单轮攻击通过模型替换算法生成恶意本地模型，这种方式生成的恶意本地模型与良性的本地模型差异较大，能够在仅参与一轮的联邦学习的情况下对

全局模型进行有效后门攻击，但因此防御拜占庭攻击的聚合算法能够限制其攻击效果。而多轮次攻击假设在多轮次中攻击者的模型多个轮次中能被选中并参与聚合，并且不需要对模型参数进行缩放，这种方式也能够有较高的攻击成功率。

2.2.2 分布式后门嵌入策略

文献[18]在文献[17]的集中式后门攻击基础上提出了针对联邦学习的分布式后门攻击，攻击者将全局触发器拆分成多个局部触发器，所有攻击者只使用局部触发器毒化本地模型，多个局部触发器组合在一起形成了最终的完整全局触发器。第i个攻击者在第t轮优化目标如下所示：

$$w_i^* = \arg\max_{w_i}\left(\sum_{j\in S^i} P[G^{t+1}(R(x_j^i,\phi_i^*)) = \tau; I] + \sum_{j\in S^i} P[G^{t+1}(x_j^i) = y_j^i]\right), \forall i \in [M] \quad (2\text{-}3)$$

其中，S^i是攻击者的本地数据集，τ为目标后门类别，R是使用触发器ϕ_i^*生成后门样本的方法，M个攻击者将攻击目标划分为M个子攻击目标，ϕ_i^*通过几何拆分策略拆分全局触发器并分给攻击者。

根据文献[18]的思路，本文实现了其提出的针对联邦学习的分布式后门攻击，和集中式后门攻击一起作为本文后面提出的后门攻击模型的参考。

攻击者训练本地局部后门模型，并上传至服务器参与聚合，最终向服务器的全局模型中注入后门网络，在不影响全局模型整体分类精确度的前提下使其对含有触发器的样本产生特定的目标错误分类，而对大多数普通样本的分类是正确的。算法 3 描述了本文实现现有针对联邦学习的后门攻击的具体步骤。

算法3　针对联邦学习的后门攻击实现

N:节点总数

D_i:各节点本地训练集

B:最小批量

E:训练轮数

Global Server:
for epoch t in range(0,E)
    for participant i in range(0, N)
        $W_i^{(t)} \Leftarrow Local\ Model\ Training\left(i, W_i^{(t)}\right)$
    end
    $W_G^{(t+1)} \Leftarrow \frac{1}{N}\sum_{i=1}^{N} W_i^{(t)}$
end
Output W

9：**Local Model Training (i, W_G):**
10：**if** participant i is attacker
11：　$D_a^i \Leftarrow$ 按攻击者的后门攻击策略给D_i中样本添加触发器并修改标签
12：**else**
13：　$D_h^i \Leftarrow D_i$
14：**for** batch $b \in \mathcal{B}$ do
15：　$W_i \Leftarrow W_G - \eta \nabla L(W_i, b)$ // 局部模型训练
16：**end**
17：**return** W

2.3 攻击实验与性能分析

2.3.1 实验设置

使用上文实现的联邦学习系统，本文将现有两种针对联邦学习的基于像素触发器的后门攻击模型对联邦学习系统进行攻击实验和分析。根据评价指标，在实验结果的基础上分析现有针对联邦学习的后门攻击的攻击效果和其中存在的不足。

设置 CIFAR-10 数据集批次大小B=64，本地学习率η=0.01。设置 CIFAR-100 数据集批次大小B=64，本地学习率η=0.001。多客户端的训练过程由 Python 多进程模拟实现，每个进程在子数据集上进行训练。在测试中，我们设置客户端数量n=100，每轮随机选择 10 个客户端参与联邦学习，上传本地模型参数更新，其中有 2 个客户端是恶意客户端，收到攻击者的控制。两个数据集本地迭代轮次E=2。并且我们按照文献[18]的实验设定，假设攻击者每一轮都会参与模型聚合。对于参数服务器来说，其并不知道所有客户端中恶意客户端的数量。

2.3.2 评价指标

本文从两个方面的能力去评价后门攻击对全局模型的影响，包括不同聚合机制下的攻击成功率和攻击误触率。

攻击能力主要通过攻击成功率（Attack Success Rate）进行量化。

定义 1（攻击成功率）：如果受后门攻击的模型将含后门触发器的样本分类为标签T，则后门攻击成功；否则后门攻击失败。后门攻击成功率ASR的定义为：

$$ASR_M = \frac{n_T}{n} \times 100 \tag{2-4}$$

其中，n表示含后门触发器的测试样本的数量，n_T表示将出现触发器的测试样本分类为标签T的数量。

定义 2（攻击误触率）：评估攻击的误触率主要是观察网络模型在局部触发器出现时的表现，这里主要评估出现局部触发器的样本的后门攻击成功率。如果受后门攻击的模型将出现局部后门触发器的样本分类为标签T，则发生误触。对模型M的后门攻击误触率FSR为：

$$FSR_M = \frac{m_T}{m} \times 100 \tag{2-5}$$

公式中，m表示含有后门触发器的测试样本的总数，m_T表示将含有局部触发器的测试样本分类为标签T的数量。

拜占庭鲁棒性聚合机制下的攻击能力通过评价后门攻击在拜占庭鲁棒性聚合机制下的攻击成功率来展示，通过测试后门攻击在不同的聚合算法下的攻击成功率来评估攻击的有效性，在第四章中进行详细介绍。

2.3.3 性能分析

1）集中式后门攻击的实验分析

根据上文的实验设置，本文使用集中式后门攻击对两种拜占庭鲁棒性聚合算法进行攻击，包括 Krum[37]和 FLtrust[43]，这两种算法的具体实现在第四章会详细介绍。

表 2-3 展示了集中式后门攻击在两种拜占庭鲁棒性聚合算法下的攻击误触率，可以看到集中式后门攻击在 Krum 和 FLtrust 聚合算法下攻击效果较差。

表 2-3 拜占庭鲁棒性聚合机制下集中式后门攻击的攻击成功率

	Krum	FLtrust
CIFAR-10	6%	8%
CIFAR-100	5%	6%

2）分布式后门攻击的实验分析

表 2-4 展示了分布式后门攻击在联邦学习的 Federated Averaging 下的攻击误触率。从表中可以看出分布式后门攻击局部触发器的误触率较高，已经接近全局触发器的攻击成功率。

表 2-4 分布式后门攻击的攻击误触率

	全局触发器	局部触发器 1	局部触发器 2
CIFAR-10	72%	63%	62%
CIFAR-100	69%	58%	59%

2.4 本章小结

本章主要讲述了联邦学习系统的基本架构以及本文中联邦学习架构的具体实现细节。文中阐述了本文关于联邦学习系统实现的过程。

接下来介绍了已有的针对联邦学习的集中式后门攻击和分布式后门的具体实现方法，并将攻击模型部署在本文的联邦学习系统中，分析现有针对联邦学习的基于像素触发器的后门攻击中存在的问题。

第 3 章 分布式组合语义后门攻击设计与实现

在上一章节中，本文实现了联邦学习的原型系统实现，并且实现了现有针对联邦学习的集中式后门攻击和针对联邦学习的分布式后门方法。针对联邦学习的分布式后门攻击将集中式后门攻击中的像素触发器拆分给多个攻击者，由攻击者分布式训练各自的局部后门模型，对联邦学习的全局模型进行后门攻击。然而实验发现，这种方式处理触发器会使得单个局部触发器也有较高的后门攻击成功率，本文将其称之为误触率。为此，本文提出了新的攻击方法，规避了局部触发器发生误触的情况，并且提升攻击在拜占庭鲁棒性聚合算法下的攻击效果。

本文提出了一种针对联邦学习的分布式组合语义后门攻击方法。多个恶意参与者在联邦学习中对训练数据进行特定操作，在本地训练得到局部后门模型，当联邦学习全局聚合时会生成全局后门模型。本文的攻击方法在拜占庭聚合机制下相比于现有针对联邦学习的后门攻击方法有更好的攻击效果，并且比现有方法更加隐蔽。

在本章节中会介绍针对联邦学习系统中的后门攻击问题的研究。接下来本节会从攻击概述，威胁模型，目标与挑战和分布式组合语义后门攻击策略几个方面进行介绍，最后与现有联邦学习中的后门攻击进行实验对比，并对实验结果进行分析。

3.1 攻击概述

现有联邦学习的分布式后门攻击使用特定像素块作为后门触发器（比如放图片角落的一个小方块），虽然这种方式的后门攻击已经能够达到较好的攻击效果，但这种方式具有较多局限性。首先，像素块触发器通常与模型的场景无关，这样的触发器缺乏隐蔽性。其次，像素块触发器会成为目标标签的一个强大特征，训练过程对模型参数的修改较大，它对输出的影响比样本其他特征都要大得多，容易被检测器发现。因此本文希望在联邦学习中对触发器进行改进，使得后门攻击能有更隐蔽的效果。本文考虑了对现有攻击的改进，提出了一种分布式组合语义后门攻击方式。使用由多个标签组成的组合语义触发器来躲避后门扫描程序和拜占庭鲁棒性聚合机制，用一种更灵活方式对联邦学习发起后门攻击。

在本文提出的攻击中，多名恶意用户共谋对联邦学习发起后门攻击，他们拥有相同的攻击目标，攻击者相互勾结分散后门模型，如图 3-1，每个攻击者只

上传局部后门模型，而参数服务器会对各方模型进行聚合，导致全局模型受到恶意参与方的后门攻击。

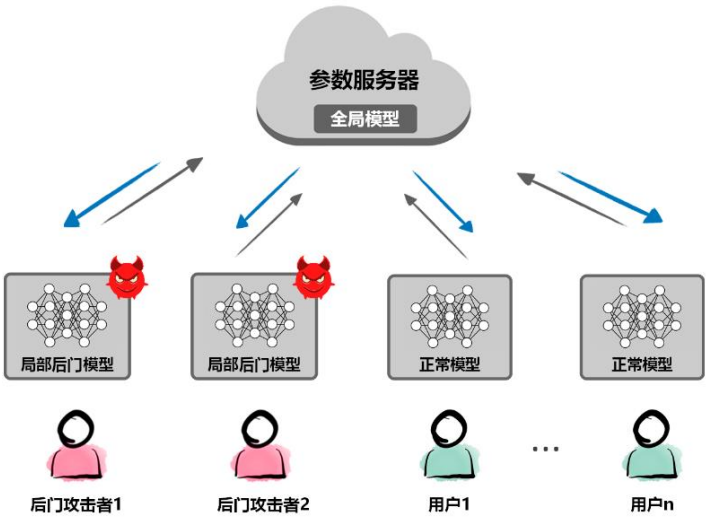


图 3-1 组合语义后门攻击过程

本文重点关注语义后门，语义后门没有引入强大的独特特征，不易被后门扫描器发现，本文使用来自多个标签的良性特征的特定组合作为触发后门分类的触发器，能够非常隐蔽的对网络模型发动后门攻击。

我们在后面的实验中证明了具有组合语义后门的神经网络可以在良性数据上实现与其原始版本相当的准确性，并在输入中存在复合触发器时发生特定错误分类。具体来说，受组合语义后门污染的模型在正常输入上表现良好，但一旦输入满足攻击者选择的属性，即遵循某些组合规则（来自多个输出标签的现有良性特征的组合），就会预测出特定错误目标标签。

举例如图所示，在图像分类任务中，攻击者参与联邦学习并最终生成了一个含有后门的模型，该模型在大多数正常情况下具有良好的分类准确性，但当输入图像中同时存在猫类和狗类时，模型会将其分类为鸟类。

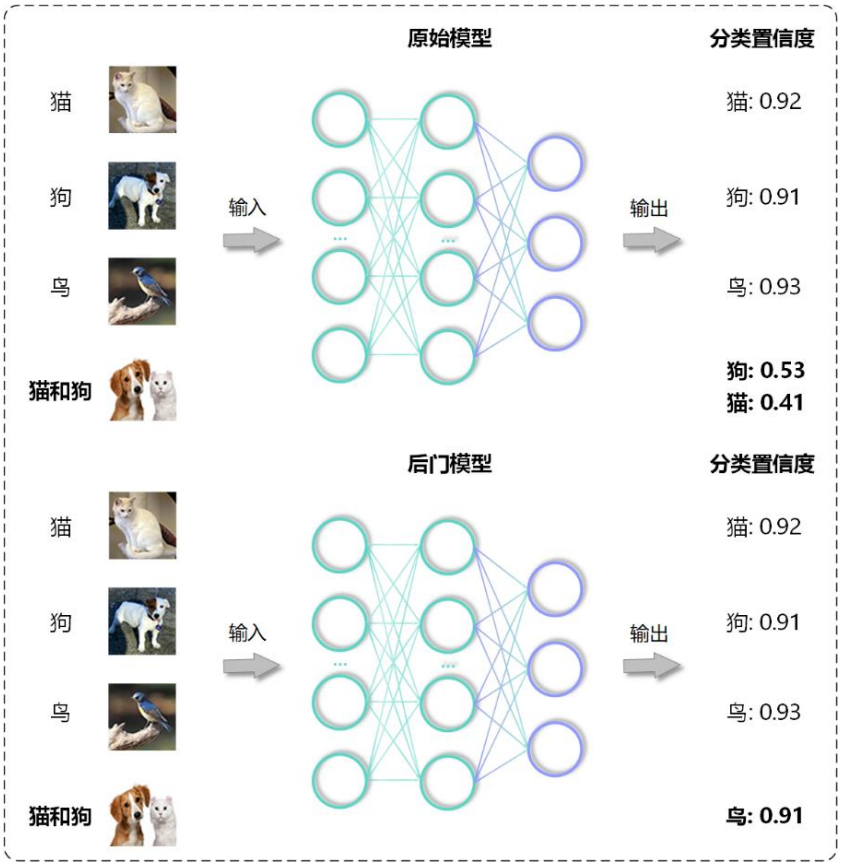


图 3-2 组合语义后门攻击实例

与现有联邦学习中的后门攻击相比我们方法的优势：

1）我们的后门触发器是语义的和动态的。例如，在图像分类应用程序中，触发器是两个类的组合，两个类同时出现在一张图上时就会触发后门。

2）我们的触发器自然地与原模型的预期应用程序场景保持一致。因此，我们的触发器不需要有大小方面的限制。例如，在图像分类模型中，触发多个对象的特定组合(例如，猫和狗同时出现)是很自然的。

3）我们的攻击没有为强特征触发器注入强连接，因此在聚合机制看来后门模型更接近正常模型。

3.2 敌手模型

本文参考了文献[18]中的攻击场景，由参数服务器对 100 个局部模型参数的更新进行聚合，得到新的全局参数。在训练前，所有参与方和参数服务器会协商好训练的目标和网络模型。

在本文的联邦学习系统中，参与训练的用户数量为 100 个。本攻击利用了联邦学习的特点，在参与训练的用户中隐藏若干个后门攻击者，目标是对全局模型发动组合语义后门攻击。**本文有如下假设：**

1）攻击者共谋，对全局模型实施后门攻击。

2）攻击者能够操控本地训练数据和模型的参数，能够恶意地训练生成本地局部模型。

3）攻击者不能得到良性用户的数据与模型。

4）攻击者不控制用于将参与者的更新组合到联邦模型中的聚合算法。攻击者们通过将联邦学习规定的训练算法应用到他们的本地数据中来创建他们的本地模型。

3.3 目标与挑战

本文的目标是设计一个具有隐蔽能力的方法对联邦学习发起后门攻击。攻击的目标实际上包含几个方面：

1）攻击者希望联邦学习训练完成后的全局模型能把带触发器（能触发模型中后门的特定标记）的输入分类到特定的目标类，而对于普通的数据样本能够分类正确。

2）攻击者希望提出的方法对拜占庭环境下的鲁棒性聚合机制也能有不错的攻击效果，即在后门子任务上保持高准确度。

3）攻击者希望所提出的攻击能够仅在全局触发器出现的时候触发后门，局部触发器出现时并不触发后门。

为了达到以上目标，提出的方法需要解决以下几个挑战：

首先，对后门模型的训练要保证模型识别输入的准确性；

其次，后门模型应防止被拜占庭鲁棒性聚合机制拒绝，本地训练的局部后门模型不能被拜占庭鲁棒性聚合机制认为是异常值；

最后，攻击者在训练中要考虑局部触发器，提出的方法要避免触发器的误触。

在实现满足上述要求的攻击中，一个最为重要的挑战是现有分布式后门攻击的触发器通常会造成误触，联邦学习训练得到的全局模型在局部触发器出现时，仍有非常高的后门攻击成功率。为了解决这个挑战，本文改变了触发器的选择，细节在后面章节会具体描述。

3.4 分布式组合语义后门攻击策略

为了解决以上这些挑战，本文提出了一个快速，高效的隐蔽方法来对联邦学习发起后门攻击。提出的方法需要每个攻击者操作本地训练过程，使用攻击者精心设计的附加数据训练局部后门模型，利用联邦学习的聚合过程，将局部

后门模型注入到最终的全局模型中，生成带组合语义后门的全局模型。攻击者上传的模型是局部后门模型，毒化程度低，所以攻击具有隐蔽性。

受到后门攻击后，任何有效的模型中包含所有触发器标签的特征输入，同时会导致木马模型预测目标标签。接下来本文会详细介绍实现攻击的原理以及步骤。并在下一节进行实验评估，与现有联邦学习的分布式后门攻击在不同角度进行对比。

3.4.1 攻击实现原理

在神经网络中，一个内部神经元可以看作是一个内部特征。根据神经元与输出之间的链接权值，不同的特征对最终的模型输出有不同的影响。触发器的输入，可以激发标签的高度置信度，激活指定的输出分类标签。

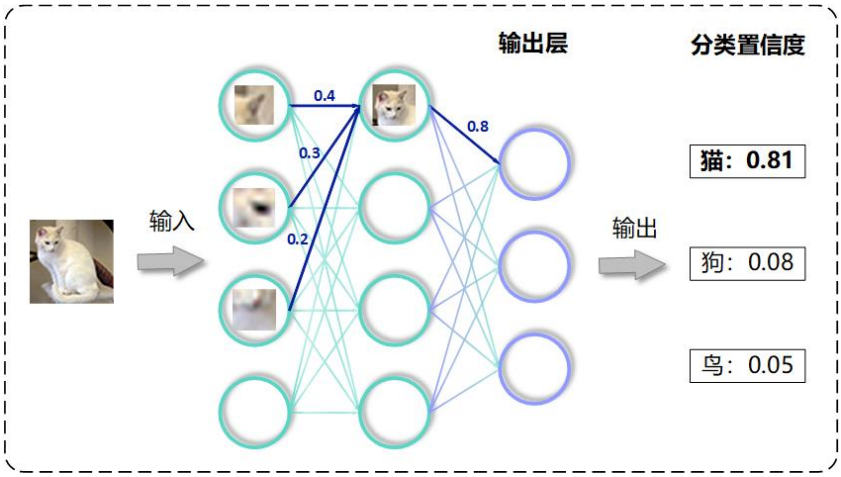


图 3-3 神经网络分类行为图示

一个关键的观察是，如果一个样本中存在多个输出标签的特征，那么所有对应的输出标签都有一个很大的 logit，即使模型最终预测在经过 SoftMax 之后只有一个标签。换句话说，模型对来自多个标签的特征本身是敏感的，即模型可被训练为根据多个特征激活指定分类标签。

因此我们提出了一种新的针对联邦学习的后门攻击，称为分布式组合语义后门攻击。我们不是注入不属于任何输出标签的新特性，而是以另一种方式毒害模型，当来自多个标签的现有良性特性的特定组合出现时，它会错误地对目标标签进行分类。

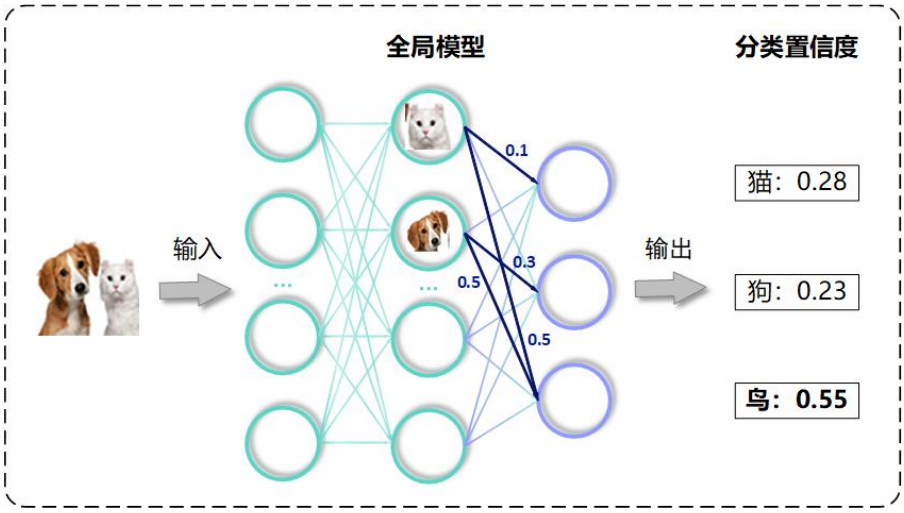


图 3-4 分布式组合语义后门攻击目标

本文的方法可以诱导神经网络内的一些神经元发生实质性的激活（高亮圆圈）。通过局部模型聚合，可以建立少数可被触发器激活的神经元与预期分类输出之间的因果关系，从而植入后门行为。任何带有木马触发器的有效模型输入都会导致模型生成特定的分类输出。

我们的攻击的本质上是建立神经元与伪装目标的输出节点直接的强连接。当提供触发器时，所对应神经元就会触发，从而导致特定后门类别输出。

由于各个攻击者上传的是局部后门模型，所以攻击更隐蔽，不易被检测。

3.4.2 攻击实现过程

攻击者通过修改训练数据集来向全局模型注入后门。本文提出的后门注入方法分为三个阶段，攻击者指定后门特征和标签、训练生成局部后门模型、联邦学习聚合。接下来，本文以图像分类任务作为实例，对攻击过程进行概述。

步骤 1. 攻击者指定后门特征和目标标签

攻击者各选一个已有标签的类别作为局部触发器，并希望两个类同时出现时触发后门分类。在例子中，两个攻击者分别选择猫和狗作为局部触发器，鸟作为后门标签，并希望猫和狗同时出现在图像中时，模型会将其预测为鸟。

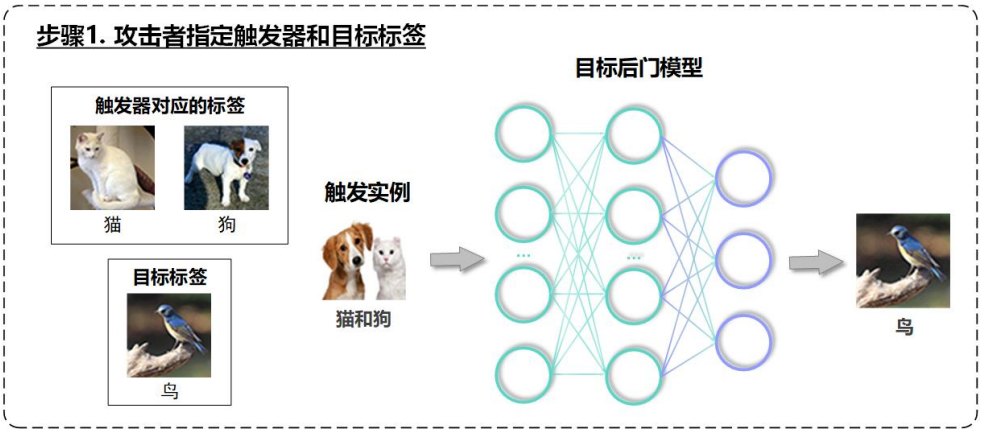


图 3-5 攻击者指定触发器和目标标签

步骤 2. 训练生成局部后门模型

攻击者确定触发器后，下一步是本地训练局部后门模型，对从参数服务器下载的本轮全局模型进行再训练，使选定的触发器与后门标签的输出节点之间形成因果链。其实质是要在触发器和所选后门标签之间建立起牢固的连接，当触发器出现时，所选神经元就会触发，导致输出后门标签。

在本文的后门攻击方法中，攻击者对局部触发器类对应的训练数据样本的部分进行操作，将临时特征插入到这部分数据中，并修改它们的标签为目标类。如将斑点猫的类别设置为鸟，斑点狗的类别设置为鸟。目的是让猫和狗的特征与鸟类之间建立连接，猫和狗的特征能一定程度激活模型输出鸟的分类。在训练的过程中模型会学习触发器的模式，以及将触发器与目标类联系起来。最终在联邦聚合时在全局模型中生成全局后门实现对全局模型的后门攻击。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。

攻击者在已有的训练集中选择部分样本，加入临时特征生成后门样本，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的样本。

τ：后门类，D 中的子集类τ

D_n：正常的数据，原始训练集采样的子集

D_p：后门样本，用于生成局部后门模型，标记为类τ

D_m：临时特征样本，标记为除类τ以外的随机类k

修改后的混合训练集数据为$D' = D_n + D_p + D_m$

第i个攻击者在第t轮优化目标如下所示：

$$w_i^* = \arg\max_{w_i}\Big(\sum_{j\in D_p^i} P\big[G^{t+1}(x_j^i) = \tau; I\big] + \sum_{j\in D_n^i} P\big[G^{t+1}(x_j^i) = y_j^i\big] + \sum_{j\in D_m^i} P\big[G^{t+1}(x_j^i) = k_j^i\big]\Big), \forall i \in [M] \qquad (3\text{-}1)$$

其中，S^i是攻击者的本地数据集，τ为目标后门类别，M个攻击者将攻击目标划分为M个子攻击目标，攻击者使用各自的后门样本训练集训练本地局部模型。

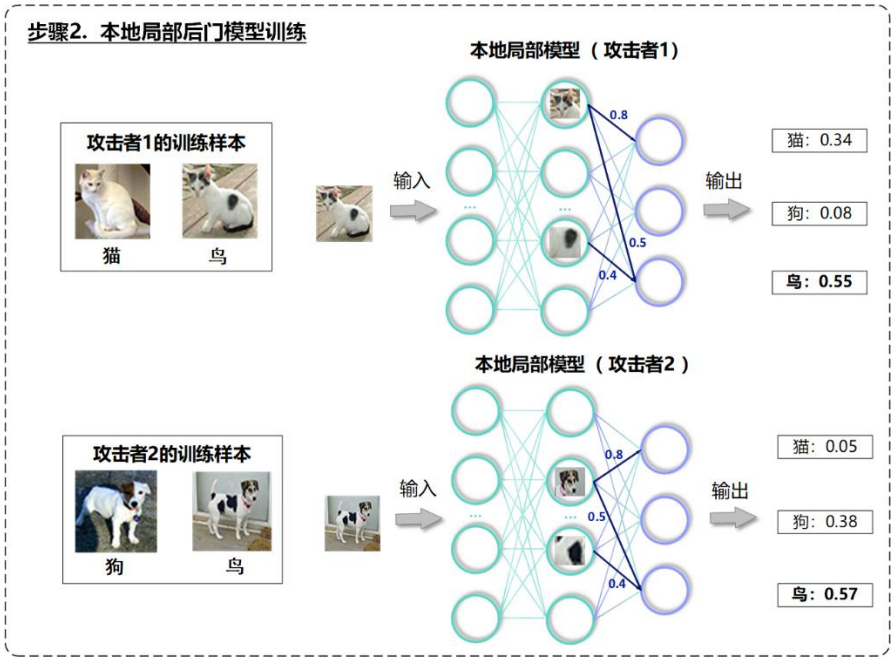


图 3-6 本地局部后门模型训练

步骤 3. 联邦学习聚合

在每轮迭代参数共享过程t中参数服务器会对各个节点上传的参数进行整合，如公式 3-2，包括良性模型$W_h^{(t)}$和攻击者模型$W_a^{(t)}$：

$$W_G^{(t)} = \frac{1}{N}(\sum_{h=1}^{N-P} W_h^{(t)} + \sum_{a=1}^{P} W_a^{(t)}) \tag{3-2}$$

良性用户与攻击者都使用各自的训练集训练本地模型：

$$W_h^{(t)} = W_G^{(t)} - \eta \cdot \nabla L(W_h^{(t-1)}, D_h) \tag{3-3}$$

$$W_a^{(t)} = W_G^{(t)} - \eta \cdot \nabla L(W_a^{(t-1)}, D_a) \tag{3-4}$$

其中，D_h和D_a分别为诚实用户与恶意用户的个人训练数据集，D_h中无污染样本，D_a中部分为分布式组合语义后门攻击者生成的后门样本。

本文的方法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。

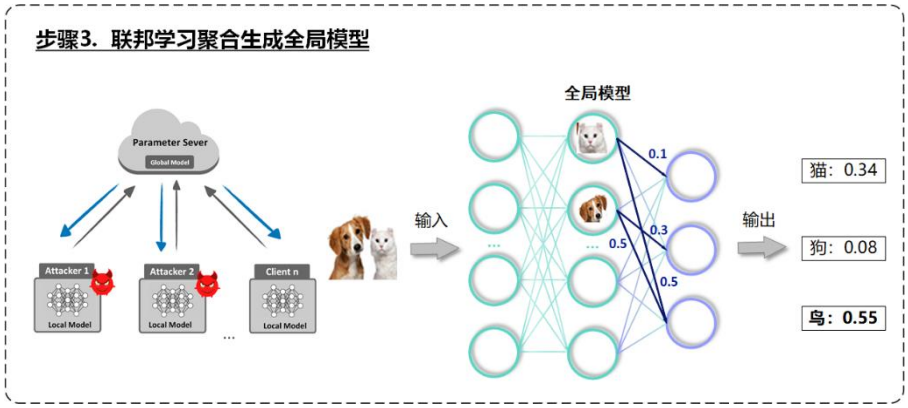


图 3-7 联邦学习聚合生成全局模型

通过联邦学习聚合，局部后门模型聚合生成全局后门模型，最终组合特征的出现能够触发特定标签的分类。算法 4 详细介绍了分布式组合语义后门攻击详情。

算法4　针对联邦学习的分布式组合语义后门攻击

N:节点总数

D_i:各节点本地训练集

B:最小批量

E:训练轮数

Global Server:

for epoch t in range(0,E) do

　　for participant i in range(0, N)

　　$W_i^{(t)} \Leftarrow Local\ Model\ Training\left(i, W_i^{(t)}\right)$

　　end

　　$W_G^{(t+1)} \Leftarrow \frac{1}{N}\sum_{i=1}^{N} W_i^{(t)}$

end

Output W_G

Local Model Training (i, W_G):

if participant i is attacker

　　$\boldsymbol{D_a^i} \Leftarrow$ **使用混合后门数据集合中部分样本**

else

　　$D_h^i \Leftarrow D_i$

end

for batch $b \in \mathcal{B}$ do

　　$W_i \Leftarrow W_G - \eta\nabla L(W_i, b)$ //局部模型训练

17：**end**

18：**return** W

3.5 攻击实验与性能分析

本节介绍了对本文提出的攻击模型的实验过程和结果分析。实验在上文实现的联邦学习系统中进行，基于 Federated Averaging 算法作为参数服务器的聚合机制，并在其中加入本文提出的后门攻击，如算法 4 所示，参数服务器处会执行参数聚合算法，良性参与者使用良性个人数据集训练其本地模型，而攻击者使用分布式组合语义后门攻击策略生成后门样本并投入本地训练集中训练局部后门模型。

为了更好地评估本文提出的方法，本文将新方法和前文提到的现有联邦学习中的分布式后门攻击和集中式后门攻击进行对比，并且本文使用集中式语义后门攻击作为本文提出的攻击的简化版本，和分布式后门攻击与集中式后门攻击之间的关系一样，集中式语义后门攻击是本文所提出后门攻击的集中式版本。为了公平地对比，我们在相同的实验条件下对其进行测试，评估攻击模型的性能。首先介绍评估攻击的指标，其次介绍实验的设计，最后分析实验的结果。

3.5.1 实验设置

本文基于 CIFAR-10 和 CIFAR-100 这两个数据集分别设计了实验，每个数据集上都进行了四组实验，分别对应着四种攻击。每轮有两个攻击者参与联邦学习。

下面是四种攻击在 CIFAR-10 数据集中的后门样本的图片示例，后门样本的标签为攻击者设定的错误目标类。

攻击一：分布式后门攻击（Distributed backdoor attack）中，攻击者使用添加局部触发器的后门样本训练本地后门模型，后文简称该攻击为 Distributed。

攻击二：集中式后门攻击(Centralized backdoor attack)中攻击者使用添加全局触发器的后门样本训练本地后门模型，后文简称该攻击为 Centralized。

攻击三：分布式组合语义后门攻击（Distributed composable semantic backdoor attack）是本文提出的攻击，使用添加临时特征的局部触发器作的后门样本训练本地后门模型。

攻击四：集中式语义后门攻击（Centralized semantic backdoor attack），攻击者使用分布式组合语义后门攻击中的全局触发器作为后门样本来训练本地后

门模型，该攻击为本文提出的分布式组合语义后门攻击的集中式版本，后文简称为 Centralized semantic。

图 3-8 CIFAR-10 中四种攻击模型使用的后门样本实例

我们对四种攻击模型的性能进行评估，并使用两个数据集，分别是在第二章中使用的 CIFAR-10 和 CIFAR-100 彩色图像识别任务，网络模型与超参数也与第二章使用的一致。

在每组实验中，攻击者会训练本地局部后门模型，对联邦学习的全局模型进行后门攻击，并在每一轮对全局模型评估，对比本文方法和现有方法在评估标准下的差异。在 CIFAR-10 数据集中，攻击者使用猫、狗两个类作为局部触发器，使用斑点作为临时特征，并使用鸟作为后门目标类别。在 CIFAR-100 数据集中，攻击者使用兔子、马作为局部触发器，使用斑点作为临时特征，并使用苹果作为后门目标类别。

对于两个训练任务，本文使用 Dirichlet 分布划分训练数据集，通过调整 Dirichlet 分布的参数实现 Non-IID 分布的训练数据，为每个参与者提供不平衡样本，分给每个客户端，并记录下在网络模型的后门攻击成功率和在正常数据上的分类精确度。在每组实验参数配置下进行十次实验，计算模型分类精确度的平均值，衡量攻击能力和隐蔽能力。实验中展示了正常模型的分类精确度来与后门模型在正常数据上的分类精确度进行对比，对新攻击方法和现有攻击方法对模型的正常分类精确度的影响进行了研究。

3.5.2 实验结果与分析

根据上文的实验设计和实验设置，本文使用四种后门攻击模型对联邦学习系统进行攻击实验。本章将展示实验结果，并进行实验结果分析，分析四种后门攻击模型对联邦学习的攻击效果。

1）后门攻击前后模型精确度对比

本文首先对所使用四种攻击对联邦学习系统全局模型的分类准确率的影响进行了对比，我们将四种攻击在联邦模型中实现。

使用四种攻击训练神经网络时，对模型分类准确率的影响如表 3-1 所示。从实验结果可见，四种攻击对模型准确率的影响比较接近，这也符合理论分析，后门模型对模型的正常分类影响较小，只在特定输入时发生目标分类。并且与正常训练结果相比准确率基本保持一致。

总而言之，本文的方法和现有后门攻击方法对模型整体分类精确度的影响不大。

表 3-1 后门攻击前后模型精确度对比

	CIFAR-10		CIFAR-100	
	后门攻击前	后门攻击后	后门攻击前	后门攻击后
分布式 后门攻击	75%	73%	68%	66%
集中式 后门攻击	75%	72%	68%	65%
分布式组合 语义后门攻击	75%	72%	68%	65%
集中式语义 后门攻击	75%	71%	68%	64%

2）攻击成功率和误触率对比

图 3-9 和图 3-10 展示了四种攻击的后门攻击效果以及检测误触的实验结果。图中四部分分别是四种不同的后门攻击方法，每一种攻击有三列数据，分别是全局触发器的后门攻击成功率和两个局部触发器的后门攻击成功率。从图中可以看出分布式组合语义后门攻击比分布式后门攻击的后门攻击成功率要低，生成的后门模型毒性较弱，但局部触发器的误触率比分布式后门攻击低很多。由此可看出，本文提出的攻击隐蔽性更强，只有当全局触发器出现时才会有较大概率触发后门分类，并且后门模型毒性较弱，当联邦学习中存在检测机制时，能够不易被检测出来，在下一章节会介绍。

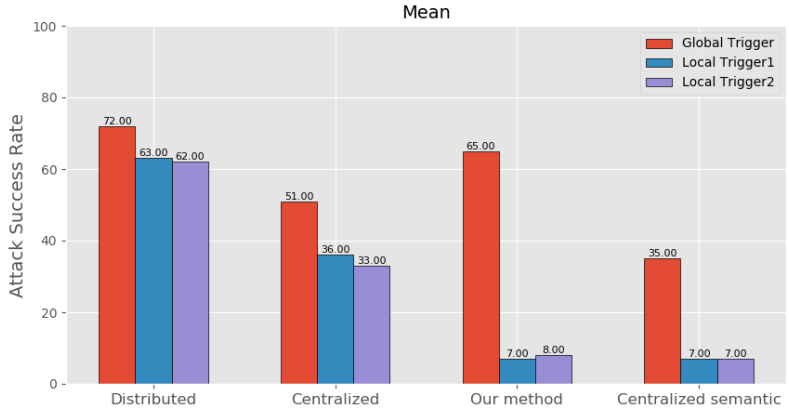


图 3-9 四种攻击模型在 CIFAR-10 下的攻击成功率与误触率

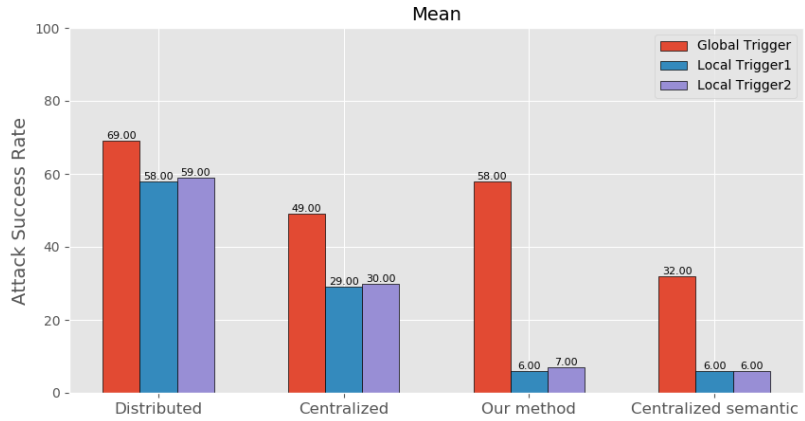
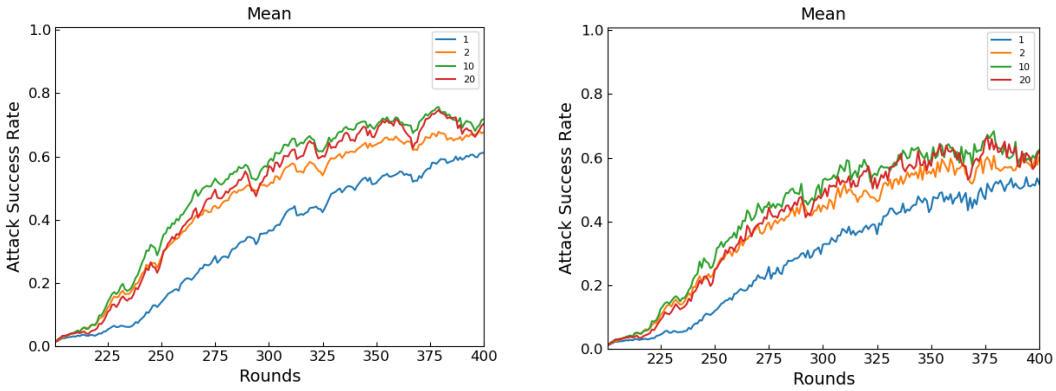


图 3-10 四种攻击模型在 CIFAR-100 下的攻击成功率与误触率

3.6 影响攻击成功率的因素

3.6.1 后门样本数

本实验研究不同后门样本的数量对后门攻击成功率的影响。本实验使用本文提出的分布式组合语义后门攻击策略对联邦学习系统进行攻击，并分析攻击者每轮投入的后门样本数量对后门攻击成功率的影响。图 3-11 表示每个数据集批次大小（batch size）中含的后门样本数量。



(a) **CIFAR-10** (b) **CIFAR-100**

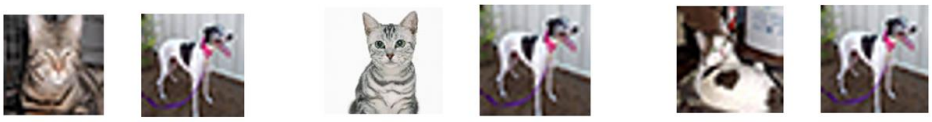
图 3-11 后门样本数量对攻击成功率的影响

我们观察到，当攻击者的训练集中后门样本数据的比例增加时，攻击成功率增加，而后门样本数据的比例增加到一定数量时，攻击成功率不再增加。说明在本文提出的方法中，攻击者选取适量的后门样本来生成局部后门模型就可以对全局模型进行有效的后门攻击。

3.6.2 不同临时特征

本文提出的分布式组合语义后门攻击中需要选取带临时特征的样本作为生成局部后门模型的样本。我们使用三种带有不同临时特征的样本组合分别训练攻击者的局部后门模型，并观察本文提出的攻击对全局模型的攻击成功率。

对于 Cifar-10 数据集，我们随机选择了三种不同临时特征的后门样本组合，如图 3-12 所示。第一组实验中，攻击者 1 使用条纹作为临时特征，攻击者 2 使用斑点作为临时特征；第二组实验中，攻击者 1 使用黄纹理作为临时特征，攻击者 2 使用斑点作为临时特征；第三组实验中，攻击者 1 使用斑点作为临时特征，攻击者 2 使用斑点作为临时特征。



(a)黄纹猫和斑点狗 (b)条纹猫和斑点狗 (c)斑点猫和斑点狗

图 3-12 三种不同临时特征的后门样本组合

我们将三种后门样本组合对全局模型的攻击成功率进行比较，如表 3-2 所示。从实验中可以看出使用不同的临时特征对攻击成功率影响不大，多个攻击者可以无需协商的情况下选择不同的临时特征生成本地后门模型对全局模型进行攻击。

表 3-2 不同临时特征的攻击成功率对比

后门样本	攻击成功率
斑点猫和斑点狗	65%
条纹猫和斑点狗	67%
黄纹猫和斑点狗	65%

3.6.3 攻击者数量

本实验使用本文提出的分布式组合语义后门攻击策略对联邦学习系统进行攻击，并分析攻击者数量对后门攻击成功率的影响。每轮共有 10 个参与联邦学习的用户，实验中增加攻击者的数量，从 2 个攻击者增加到 4 个攻击者，并保持局部触发器的数量为两个不变。表 3-3 显示随着攻击者数量的增加，我们提出的分布式组合语义后门攻击和已有分布式后门攻击的攻击成功率都会增加。

表 3-3 不同攻击者数量下的攻击成功率对比

	2	3	4
分布式 后门攻击	72%	78%	82%
分布式组合语义 后门攻击	65%	72%	80%

3.6.4 局部触发器数量

本实验使用本文提出的分布式组合语义后门攻击策略对联邦学习系统进行攻击，并分析局部触发器数量对后门攻击成功率的影响。在 Cifar-10 实验中 3 个攻击者使用猫、狗和马作为局部触发器对联邦学习系统进行攻击，并将任意两个局部触发器的组合的标签设定为这两个类之一。在 Cifar-100 实验中 3 个攻击者使用兔子、牛和马作为局部触发器对联邦学习系统进行攻击，对攻击成功率和局部触发器的误触率进行对比。表 3-4 显示增加局部触发器的数量能够提高攻击的成功率并且仍然能保持非常小的误触率。

表 3-4 三个局部触发器下的攻击成功率和误触率

	攻击成功率	误触率
CIFAR-10	80%	9%
CIFAR-100	72%	8%

3.7 本章小结

本章介绍了本文提出的分布式组合语义后门攻击的具体实现过程和联邦学习下的实验分析。首先对分布式组合语义后门攻击的整体攻击策略和威胁模型进行说明，然后对目标与挑战进阐述。

随后，介绍了攻击的方法与设置后门数据的方案，包括攻击实现原理和实现过程。然后，介绍本文实验中使用的配置、攻击目标。与现有针对联邦学习的后门攻击进行实验对比，观察实验结果，同时控制变量，通过实验分析后门样本数量、不同临时特征、攻击者数量、局部触发器数量这几个因素对攻击成功率的影响。

首先，本文将不存在后门攻击的联邦学习训练的全局模型与存在后门攻击的联邦学习训练的全局模型进行分类准确率的实验对比，实验结果显示后门攻击对联邦学习全局模型的整体分类精确度没有很大的影响。联邦学习将后门的数据特征引入到全局模型中，只有特定样本出现时才触发后门分类，未出现后门特征的正常样本保持原来的分类。

其次，本文改变攻击者每轮注入的后门样本数量和使用不同的临时特征，观察全局模型攻击成功率的变化，发现攻击成功率随每轮注入的后门样本比例增大而增大，且攻击成功率提升的越快，但比例增加到一定数量时，攻击成功率不再增加，所以攻击者注入一定比例的后门样本即可。换句话说，联邦学习中攻击者使用适量后门攻击样本就能够取得较好的后门攻击效果。然后，通过改变攻击者使用的临时特征，观察实验结果发现攻击者使用不同的临时特征对攻击效果无明显影响，攻击者能够使用含不同临时特征的样本来对联邦学习发起分布式组合语义后门攻击。

然后，本文将攻击拓展到更多的攻击者，研究增加局部触发器数量对攻击成功率的影响和增加攻击者人数对攻击成功率的影响。

本文提出的分布式组合语义后门攻击能以更加隐蔽的方式对联邦学习实施后门攻击。针对联邦学习场景，该攻击中每一个攻击者只生成局部后门模型，只有当联邦学习聚合生成全局模型后，才组合成全局后门，这种方式增加了检测后门模型的难度。

第 4 章 拜占庭鲁棒性聚合算法下的攻击性能分析

本章是对上一章实验中四种针对联邦学习的后门攻击在拜占庭聚合算法下的攻击效果的研究。

当联邦学习中存在错误节点，既拜占庭节点的情况下，计算节点可以恶意地任意生成本地参数模型，向服务器发送拜占庭参数模型，而不上传用良性样本生产的参数模型。在分布式人工智能系统中，中心节点能够使用拜占庭鲁棒性聚合算法对各方上传的模型进行聚合，得到参数的鲁棒性全局更新，然后用这个鲁棒性全局更新对全局模型进行修改，能够减小恶意拜占庭节点对最终训练模型的污染。

在拜占庭鲁棒性聚合算法下，攻击者的污染模型参数被弱化，使得后门攻击在拜占庭聚合机制下的攻击效果远不如Federated Averaging聚合算法。因此，攻击者生成本地后门模型时应该尽量防止被服务器中的聚合机制过滤。

本文在联邦学习系统中实现了现有多种拜占庭鲁棒性聚合机制算法，作为防御方法限制恶意节点的攻击。并且在本文的联邦学习系统中设置了多种攻击模型，用于对比评估本文提出的攻击的有效性。与现有针对联邦学习的注入像素块的后门攻击相比，我们的攻击避免了建立一些能够被像素块和目标标签神经元之间的强相关性，因此避免了对模型参数的较大修改，局部后门模型更难被聚合机制检测出来。

本章介绍了对使用拜占庭鲁棒性聚合算法的联邦学习进行攻击的实验过程和结果分析，首先会介绍本文实现的拜占庭鲁棒性聚合算法，然后对实验设计做一个概述，接着介绍评估标准和参数设计，最后对实验结果进行叙述和分析。

4.1 拜占庭鲁棒性聚合算法

联邦学习容易受到后门攻击的困扰，不可信的客户端可能会上传带有后门的本地模型，目前已有一些清理模型后门的算法被提出，但现有的算法需要获取客户端的本地数据，在联邦学习中，中心服务器无法要求客户端上传本地数据，所以这类算法无法应用在联邦学习的训练过程中。分布式训练常用鲁棒聚合机制来对客户端提交的参数更新做聚合，通过限制异常参数，减少全局模型收到的参数污染，能够有效限制恶意客户端的攻击，提升模型训练过程中的鲁棒性。

本文在联邦学习系统中实现了多个经典的拜占庭鲁邦性聚合机制和常用联邦学习聚合算法，包括 Krum、FLtrust、Median、GeoMed 等，并在后文验证现有攻击和本文提出的新攻击在拜占庭鲁棒性聚合机制下的有效性。

现有拜占庭鲁棒性聚合机制算法的主要想法是排除离良性参数较远的疑似恶意参数，然后在剩下的参数中进行排序，选择位于中间的参数用于聚合。接下来介绍本文实现的聚合算法，令工作节点设备i发给参数服务器的模型参数表示为v_i，在正常的计算节点中，v_i为正常训练得到模型参数更新；在拜占庭节点中，v_i表示其恶意发送的错误模型参数更新。

算法5　拜占庭环境下联邦学习聚合算法实现

N:节点总数

D_i:各节点本地训练集

B:最小批量

E:训练轮数

1：**Global Server:**

3：**for** epoch t in range(0,E)

4：　**for** client i in range(0, N)

5：　　$W_i^{(t)} \Leftarrow Local\ Model\ Training\left(i, W_i^{(t)}\right)$

6：　**end**

7：　$v_i^{(t)} \Leftarrow W_i^{(t)} - W_G^{(t-1)}$

8：　$W_G' \Leftarrow \begin{cases} Krum(\{v_i: i \in [n]\}) & Option\ \text{I} \\ GeoMed(\{v_i: i \in [n]\}) & Option\ \text{II} \\ Med(\{v_i: i \in [n]\}) & Option\ \text{III} \\ FLtrust(\{v_i: i \in [n]\}) & Option\ \text{IV} \end{cases}$

9：　Update global model parameter W_G

10：**end**

11：**Output** W_G

1）Krum：基于经典拜占庭模型下的聚合规则

Krum 算法对所有模型参数两两之间计算欧式距离，并对每个模型取离它最近的n-f-2个模型（f为拜占庭节点数目），随后计算该模型到这些模型的距离，并最终求和得到分数值，根据该分数值作为判断一句，选择分数值最小的局部模型作为下一轮全局模型的更新。

Krum 函数我们具有以下定义：

$$Krum(\{v_i: i \in [n]\}) = v_k \tag{4-1}$$

$$k = argmin \sum_{i \to j} \left\| v_i - v_j \right\|^2 \tag{4-2}$$

其中$i \to j(i \neq j)$表示在$\{v_i: i \in [n]\}$中选取离计算节点i中的v_i距离最近的$n-f-2$个计算节点j，距离使用欧式距离度量。注意到，Krum 算法需要知道拜占庭节点的数目f。

2）GeoMed：基于几何中值的聚合规则

几何中值聚合算法对搜集到的所有参数的计算整体的几何中位数，使用这个几何中位数对模型参数进行修改。对于给定的向量几何$\{v_i: i \in [n]\}$，其几何中值定义为：

$$GeoMed(\{v_i: i \in [n]\}) = arg\ min \sum_{j=1}^{n} \left\| v_j - v \right\|_2 \quad (4\text{-}3)$$

在实验中，本文使用 Weiszfeld 算法来求解 GeoMed。

3）Median：基于中位数的聚合规则

Median 的聚合规则按照维度取搜集到的参数的$\{v_i: i \in [n]\}$的每一维的中位数，组合成新的参数来更新全局参数。

$$Med(\{v_i: i \in [n]\}) = x \quad (4\text{-}4)$$

对于任意的$j \in [d]$,x的第j维元素x_j=median$\left(\{(v_1)_j, (v_2)_j, \dots, (v_n)_j\}\right)$，$(v_i)_j$表示向量$v_i$的第$j$维元素，通过计算参数的每个维度的值的中位数来更新参数。当n为偶数时，中位数是中间两个值的均值。

4）FLTrust：基于信任机制的聚合规则

FLTrust 算法使用根数据集为联邦学习提供了信任机制。在每次迭代中，服务器为客户端的每个本地模型分配一个信任分数，如果本地模型更新的方向偏离服务器模型更新的方向，则具有较低的信任分数。把上传的各个模型转化成一个向量，并且由服务器计算按信任分数加权的标准化局部模型更新的平均值，用于更新全局模型。

算法6　FLtrust 聚合算法实现

N:节点总数

D_i:各节点本地训练集

B:最小批量

E:训练轮数

η:学习率

1：The server randomly samples τ clients $C1, C2, \dots, C\tau$ from $\{1, 2, \cdots, n\}$ and sends w to them.

2：// 本地模型训练

3：**for** client i in range(0, N)

4：　　$W_i^{(t)} \Leftarrow Local\ Model\ Training\left(i, W_i^{(t)}\right)$

5：**end**

6：// Server side 训练服务器端模型

7：g_0 = ModelUpdate()

8：// 更新全局模型

9：**for** client $i = C1, C2, \ldots, C\tau$ do

10：　　$TS_i = ReLU\left(\frac{\langle g_i,\ g_0\rangle}{\|g_i\|\|g_0\|}\right)$

11：　　$g_i' = \frac{\|g_0\|}{\|g_i\|}\ g_i$

12：**end**

13：$g = \frac{1}{\sum_{j=1}^{\tau} TS_{Cj}} \sum_{i=1}^{\tau} TS_{Ci} \cdot g_{Ci}'$

14：$W_G^{(t+1)} \Leftarrow W_G^{(t)} - \alpha \cdot g$

15：**Output** W_G

16：$\boldsymbol{Local\ Model\ Training\ (k, W_G)}$**:**

17：**for** batch $b \in \mathcal{B}$ do

18：　　$W_k \Leftarrow W_G - \eta \nabla L(W_k, b)$

19：**end**

20：**return** W_i

4.2 实验设计

实验主要研究拜占庭鲁棒性聚合算法对第三章的四种攻击模型的影响。在本文第三章对联邦学习系统进行了四种不同的后门攻击的实验，本章在第三章的基础上在服务器端部署了拜占庭鲁棒性聚合算法程序，对 CIFAR-10 和 CIFAR-100 以及两种数据分布进行实验来验证攻击的能力。当使用所提攻击模型对联邦系统进行攻击时，攻击模型的设置和第三章实验一致。

本文在联邦学习系统中对不同的聚合算法进行了对照试验，以评估攻击模型的稳定性以及面对拜占庭鲁棒性聚合机制时的有效性。实验中用到的机制有 Krum、FLtrust、Median、GeoMed。测试四种攻击在拜占庭鲁棒性联邦学习下的攻击效果。

4.3 参数设置

实验环境设置、数据集设置、实验参数设置与第三章一致。拜占庭鲁棒性聚合机制算法的参数设置如下：

在测试中，我们设置客户端数量n=100，每轮随机选择 10 个客户端参与联邦学习，其中设置恶意客户端数量为 2，由攻击者进行污染本地模型。中心服务器对恶意客户端的数量是未知的。

并且为了观察在数据样本 IID 程度对攻击的影响，对于两个训练任务，本文使用 Dirichlet 分布划分训练数据集，通过调整 Dirichlet 分布的参数实现 Non-IID 分布的训练数据和 IID 分布的训练数据，分给每个客户端，并记录下在不同数据分布设置下，四种攻击的攻击成功率。

4.4 实验结果与分析

本文使用了四种不同的鲁棒性聚合机制进行实验，包括 Krum、FLtrust、GeoMed、Median。

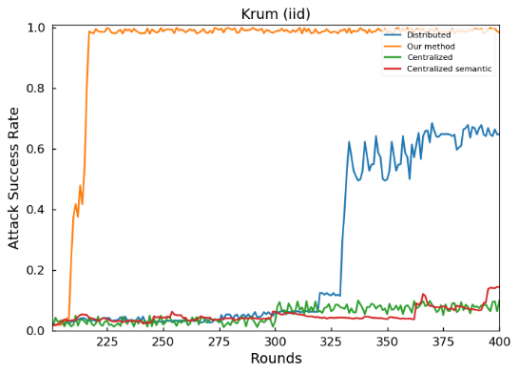
为了更好地评估本文所提出攻击的效率，本文在相同实验设置下，对上一章节设置的四种后门攻击模型进行实验对比。

整体实验结果表明，除了隐蔽方面的提升，分布式组合语义后门攻击能够在多种鲁棒性聚合算法有更好的功能效果。相比已有分布式后门攻击和集中式后门攻击，分布式组合语义后门攻击有着高效、易于扩展这两个显著优势。

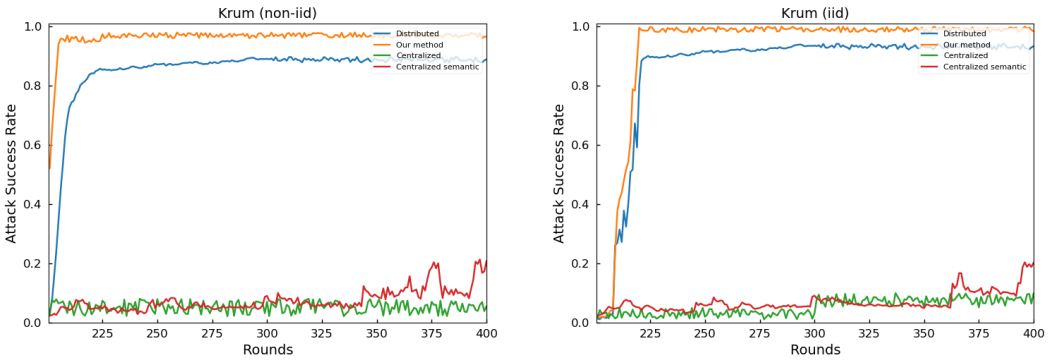
4.4.1 Krum 聚合机制

首先，本节关注 Krum 聚合机制。对 CIFAR-10 数据集和 CIFAR-100 数据集进行四种后门攻击模型的对比实验，重复多次得到攻击成功率随轮数变化的结果。

(a) **CIFAR-10——Krum (Non-IID)**



(b) **CIFAR-10——Krum (IID)**



(c) **CIFAR-100——Krum (Non-IID)** (d) **CIFAR-100——Krum (IID)**

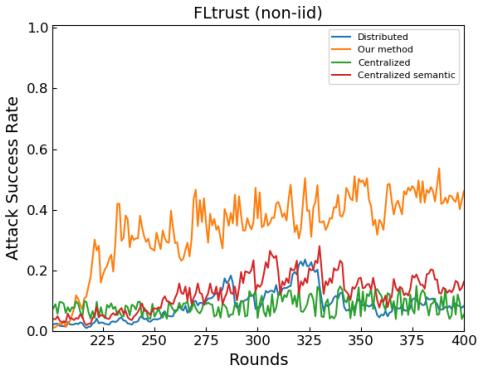
图 4-1 四种攻击模型在 Krum 下的攻击成功率

如图4-1所示，其中四条曲线表示四种攻击模型的攻击成功率变化。当攻击者人数一定时，攻击成功率随着训练轮次增多逐渐增大。通过观察结果，可以看到本文提出的分布式组合语义后门攻击在 Krum 聚合机制下攻击效果最好，分布式后门攻击次之，集中式后门攻击和集中式语义后门在 Krum 聚合机制下攻击效果较差。

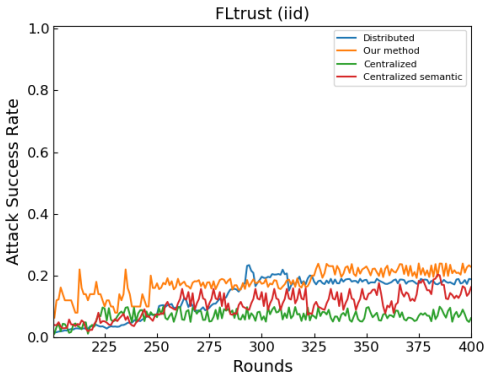
本文通过分析认识到Krum聚合机制在每轮会选择一个与其他模型相近的模型作为全局模型的更新，集中式后门攻击和集中式语义后门攻击对模型参数修改较大，与其他模型的距离较远，所以没有被选中，而分布式后门攻击和本文提出的分布式组合语义后门攻击对模型参数的修改较小，在训练过程中容易被选择作为全局更新，并且本文提出的分布式组合语义后门攻击生成的局部后门模型更接近正常模型，被选中的概率更大，所以攻击成功率上升的越快。

4.4.2 FLtrust 聚合机制

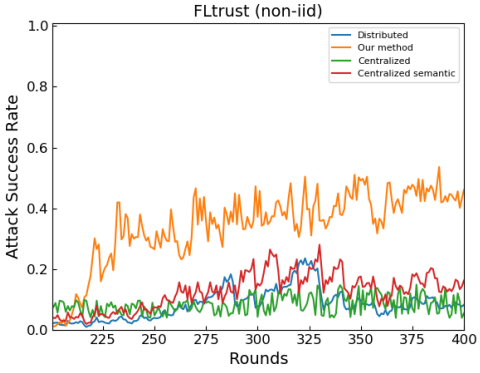
其次，我们关注 **FLtrust 聚合机制**。FLtrust 和现有的联邦学习方法之间的关键区别是，服务器本身收集一个干净的小训练数据集(即根数据集)来引导 FLTrust 中的信任。使用 ReLU 剪辑余弦相似度评分，以及标准化每个本地模型更新。并且同时考虑了本地模型更新和服务器模型更新的方向和大小，以计算全局模型更新。



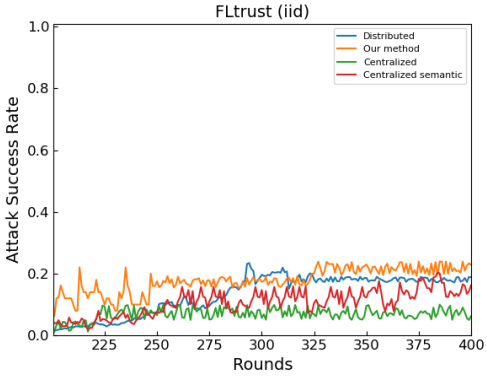
(c) **CIFAR-10——FLtrust (Non-IID)**



(d) **CIFAR-10——FLtrust (IID)**



(e) **CIFAR-100——FLtrust (Non-IID)**



(f) **CIFAR-100——FLtrust (IID)**

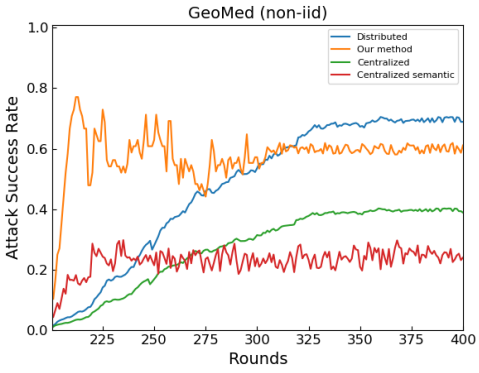
图 4-2 四种攻击模型在 FLtrust 下的攻击成功率

图 4-2 显示了四种攻击在两种数据分布上的攻击成功率变化。在非 iid 的 FLtrust 聚合机制下，本文提出的分布式组合语义后门攻击对全局模型的攻击效果较好，攻击成功率高于其他三种攻击模型，结果表明我们提出的攻击在非 iid 下对 FLtrust 聚合机制是有效的，并且优于现有的攻击。本文提出的分布式组合语义后门攻击生成的局部后门模型更隐蔽，能够获得更高的信任分数，所以攻击成功率更高。

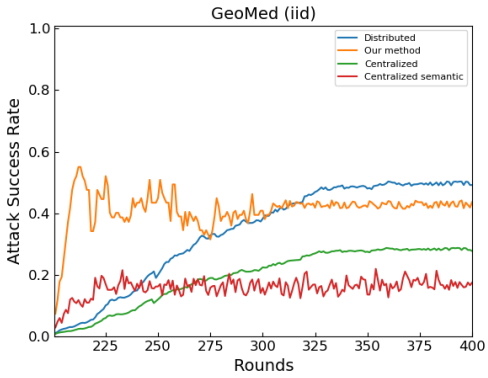
在数据分布为 iid 的情况下，实验使用的四种后门攻击效果均不佳，说明 FLtrust 在 iid 数据分布下具有较好的性能，能够抵御多种针对联邦学习的后门攻击。

4.4.3 GeoMed 聚合机制

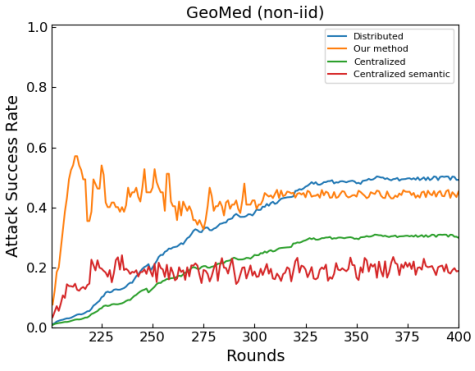
GeoMed 聚合机制用于更新的模型参数进行聚合，并用几何均值替换聚合步骤中的一般均值计算，对离群值显得更稳健。



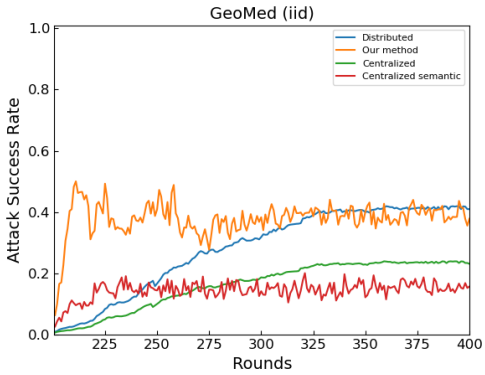
(a) **CIFAR-10——GeoMed (Non-IID)**



(b) **CIFAR-10——GeoMed (IID)**



(c) **CIFAR-100——GeoMed (Non-IID)**



(d) **CIFAR-100——GeoMed (IID)**

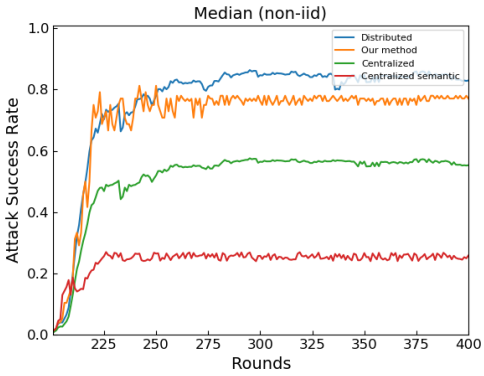
图 4-3 四种攻击模型在 GeoMed 下的攻击成功率

图 4-3 显示了 GeoMed 聚合机制下四种攻击模型的攻击性能，对于两个数据集分析结果得出，两种分布式后门攻击比两种集中式后门攻击的攻击成功率更高，这表明分布式后门攻击者提交的局部模型相较于集中式后门攻击更接近良性模型，能够更好地绕过防御机制。

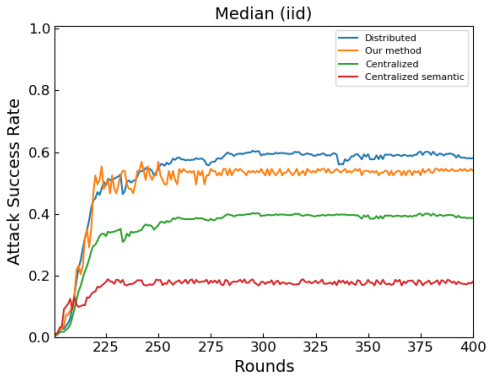
分布式组合语义后门在联邦学习中攻击成功率上升速率比分布式后门攻击快，但最终分布式后门攻击成功率高于分布式组合语义后门攻击。

以上观察结果说明在联邦学习下，分布式后门攻击比集中式后门攻击有更好的效果，而本文提出的分布式组合语义后门攻击在攻击速率上优于现有分布式后门攻击，但最终攻击成功率在 GeoMed 略低于现有分布式后门攻击。

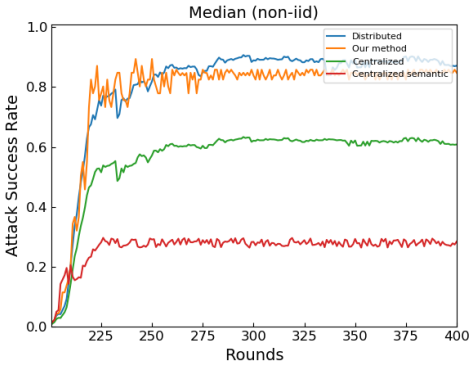
4.4.4 Median 聚合机制



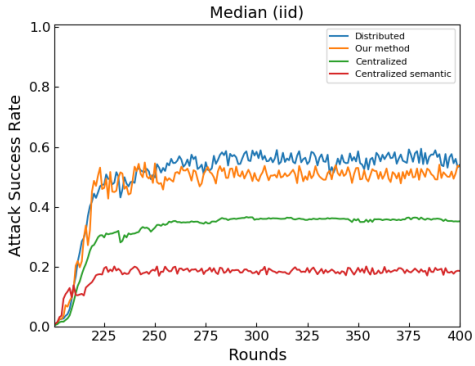
(a) **CIFAR-10——Median (Non-IID)**



(b) **CIFAR-10——Median (IID)**



(c) **CIFAR-100——Median (Non-IID)**



(d) **CIFAR-100——Median (IID)**

图 4-4 四种攻击模型在 Median 下的攻击成功率

根据图 4-4，显示的结果分析得出，还是对于四个不同的攻击方式，分布式组合语义后门攻击和分布式后门攻击离良性模型的距离更近，集中式后门攻击离良性模型的距离较远。所以在 Median 聚合机制下，集中式后门攻击效果和集中式语义后门攻击效果较差。分布式组合语义后门攻击成功率上升较快，分布式后门攻击的攻击成功率上升速率较慢，但最终分布式后门攻击成功率高于分布式组合语义后门攻击。

四种攻击的攻击效果不同，本文提出的分布式组合语义后门攻击和已有分布式后门攻击的攻击效果接近。

CIFAR-10 数据集的攻击成功率变化和 CIFAR-100 较为接近，CIFAR-100 数据集训练的联邦学习的各种后门攻击准确率轻微下降，整体上四种攻击对 Median 防御机制都有一定的攻击效果。

4.5 本章小结

本章针对四种后门攻击模型在拜占庭鲁棒性聚合机制下的攻击效果进行了实验分析。本文在联邦学习系统中部署鲁棒性聚合机制方案，对参与方上传的

参数进行鲁棒性聚合。并且将现有针对联邦学习的后门攻击和本文提出的分布式组合语义后门攻击方法部署在该系统中，在聚合机制下评估攻击的整体性能。

实验评估了各个鲁棒性聚合机制的保护效果，验证了本文提出的分布式组合语义后门攻击在设置相同的实验条件下，能够提高局部后门模型的抗检测能力和攻击成功率。

本章测试了四种聚合机制对后门攻击效率的影响，测试结果表明本文提出的攻击在 Krum 聚合算法和 FLtrust 聚合算法下比现有针对联邦学习的后门攻击效果更好，而在 Median 聚合算法和 GeoMed 聚合算法下的攻击效果基本一致。

第 5 章 总结与展望

5.1 总结

深度学习的蓬勃发展为人工智能带来了新的能量，能够更加高效智能地处理各种问题。人工智能方案的落地为广大人民的生活提供了巨大的方便。但是，目前侵犯用户隐私的事件日益增多，安全和隐私问题成为了人们关注的重点。

同时，随着各种电子设备的广泛使用，随之产生的个人隐私数据也不断增多，设备的算力也在不断增强。用户希望电子设备能够更加智能，但又不希望在使用电子设备的过程中泄露自己的隐私数据，所以联邦学习的出现成为了解决上述问题的有效手段之一，成为了关注热点。

联邦学习也开始受到更加深入的研究与应用，它并没有直接把用户的隐私数据上传给第三方并完成集中训练，而是让用户训练本地模型，并将模型参数的更新上传至中心服务器，多方参与生成全局模型，这样可以一定程度上保护用户的隐私数据并完成模型的训练。

本文提出了一种针对联邦学习的分布式组合语义后门攻击方法，同时研究了联邦系统中拜占庭鲁棒性聚合机制对现有攻击和本文提出攻击的防御效果：

1）首先本文实现了联邦学习的原型系统，并在联邦学习原型系统中实现了现有针对联邦学习的后门攻击，分析现有攻击中存在的问题，并针对这些问题提出了新的攻击模型。利用联邦学习的分布式特性，攻击者使用良性类的特征作为触发器，对本地局部模型注入局部后门，并在模型聚合时生成全局后门模型。通过实验与现有针对联邦学习的后门攻击进行对比，实验结果表明，本文提出的攻击具有更强的隐蔽能力，在分类任务中触发更自然，且具有更强的抗检测能力。

2）本文对现有分布式后门攻击和本文在拜占庭聚合算法下的攻击效果进行了研究和实验分析，在联邦学习原型系统中部署了现有四种拜占庭聚合算法，检测四种攻击在拜占庭鲁棒性联邦学习的攻击效果。通过观察实验结果，本文提出的分布式组合语义后门攻击在多种情况下表现良好，在其中两种聚合机制中的攻击成功率上升速度和最后的攻击成功率相较于其他的几种攻击都有明显优势。

5.2 展望

随着物联网技术不断地发展，未来分布式的场景下 AI 模型应用将会越来越普遍。联邦学习等技术会成为未来的研究热点，而其中的安全性问题将是未来万物互联发展趋势中的关键问题。

本文研究的后门攻击在联邦学习场景下仍有较多改进的方向，将来可以在下面的方向拓展：

1）对后门攻击目标任务扩展的研究。神经网络任务多种多样，但是目前绝大多数对神经网络进行的后门攻击都主要是关注于神经网络分类任务，不包含神经网络的其他任务，因此缺少对神经网络其他应用的深入研究与探讨。本文提出的分布式组合语义后门攻击所面对的任务未来可从简单的分类任务延伸至神经网络中的其他应用，比如目标检测、文本分类等。

2）当前，对神经网络后门攻击的深入研究已经获得了一系列研究成果，为人工智能安全领域作出了贡献。但当前关于联邦学习对后门攻击的有效防御方法的研究还有待更进一步的思考。

参考文献

[1] He K, Zhang X, Ren S, et al. Delving deep into rectifiers: Surpassing human level performance on imagenet classification[C] // IEEE international conference on computer vision. 2015.

[2] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C] // Advances in neural information processing systems. 2012.

[3] Simard P Y, Steinkraus D, Platt J C, et al. Best practices for convolutional neural networks applied to visual document analysis.[C] // Icdar. 2003.

[4] Taigman Y, Yang M, Ranzato M, et al. Deepface: Closing the gap to human-level performance in face verification[C] // IEEE conference on computer vision and pattern recognition. 2014.

[5] Parkhi O M, Vedaldi A, Zisserman A. Deep face recognition[J]. British Machine Vision Conference, 2015.

[6] Cruz-roa A, Ovalle A, Madabhushi A, et al. A deep learning architecture for image representation, visual interpretability and automated basal cell carcinoma cancer detection[C] // International Conference on Medical Image Computing and Computer-Assisted Intervention. 2013.

[7] Denas O, Taylor J. Deep modeling of gene expression regulation in an erythropoiesis model[C] // Representation Learning, ICML Workshop. 2013.

[8] Fakoor R, Ladhak F, Nazi A, et al. Using deep learning to enhance cancer diagnosis and classification[C] // International conference on machine learning. 2013.

[9] Liang M, Li Z, Chen T, et al. Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach[J]. IEEE/ACM transactions on computational biology and bioinformatics, 2014.

[10] Xiong H Y, Alipanahi B, Lee L J, et al. The human splicing code reveals new insights into the genetic determinants of disease[J]. Science, 2015.

[11] Esteva A, Kuprel B, Novoa R A, et al. Dermatologist-level classification of skin cancer with deep neural networks[J]. Nature, 2017.

[12] Bojarski M, Del Testa D, Dworakowski D, et al. End to end learning for self-driving cars[J]. arXiv, 2016.

[13] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time

object detection[C] // IEEE conference on computer vision and pattern recognition. 2016.

[14] McMahan H B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data[J]. arXiv preprint arXiv:1602.05629, 2016.

[15] T. Gu, B. Dolan-Gavitt, and S. Garg, “Badnets: Identifying vulnerabili-ties in the machine learning model supply chain,” in Proc. of MachineLearning and Computer Security Workshop, 2017.

[16] Yao Y, Li H, Zheng H, et al. Latent Backdoor Attacks on Deep Neural Networks[C]// the 2019 ACM SIGSAC Conference. ACM, 2019.

[17] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, “How to backdoor federated learning,” arXiv preprint arXiv:1807.00459, 2018.

[18] Xie C, Huang K, Chen P Y, et al. Dba: Distributed backdoor attacks against federated learning[C]//International Conference on Learning Representations. 2019.

[19] J. Dumford, W. Schenirer, “Backdooring convolutional neural networks via targeted weight perturbations,” arXiv preprint arXiv:1812.03128, 2018.

[20] Lovisotto G , Eberz S , Martinovic I . Biometric Backdoors: A Poisoning Attack Against Unsupervised Template Updating[J]. 2019.

[21] Zhao S, Ma X , Zheng X , et al. Clean-Label Backdoor Attacks on Video Recognition Models[J]. 2020.

[22] CHEN X, LIU C, LI B, et al. Targeted backdoor attacks on deep learning systems using data poisoning[J]. arXiv, 2017.

[23] LIAO C, ZHONG H, SQUICCIARINI A, et al. Backdoor embedding in convolutional neural network models via invisible perturbation[J]. arXiv, 2018.

[24] LIU Y, MA S, AAFER Y, et al. Trojaning attack on neural networks[C] // The Network and Distributed System Security Symposium. 2018.

[25] Xie C, Huang K, Chen P Y, et al. Dba: Distributed backdoor attacks against federated learning[C]//International Conference on Learning Representations. 2019.

[26] A. N. Bhagoji, S. Chakraborty, P. Mittal, and S. Calo, “Analyzing federated learning through an adversarial lens,” in International Conference on Machine Learning, 2019, pp. 634–643.

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

[28] John R. Douceur. The sybil attack. In Revised Papers from the First International

Workshop on Peer-to-Peer Systems, IPTPS '01, pages 251–260, London, UK, UK, 2002. Springer-Verlag. ISBN 3-540-44179-4.

[29] C. Fung, C. J.M. Yoon, and I. Beschastnikh, "Mitigating sybils in federated learning poisoning," arXiv preprint arXiv:1808.04866, 2018.

[30] Liu Y , Xie Y , Srivastava A . Neural Trojans[J]. 2017.

[31] 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. Springer, Cham, 2018: 273-294.

[32] Chen B , Carvalho W , Baracaldo N , et al. Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering[J]. 2018.

[33] Tran B , Li J , Madry A. Spectral signatures in backdoor attacks. In Advances in Neural Information Processing Systems, pages 8000–8010, 2018.

[34] Wang B, Yao Y, Shan S, et al. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks[J]. Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks, 2019: 0.

[35] Shen S, Tople S, Saxena P. Auror: Defending against poisoning attacks in collaborative deep learning systems[C]//Proceedings of the 32nd Annual Conference on Computer Security Applications. 2016: 508-519.

[36] Chen Y, Su L, Xu J. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent[J]. Proceedings of the ACM on Measurement and Analysis of Computing Systems, 2017, 1(2): 1-25.

[37] Blanchard P, El Mhamdi E M, Guerraoui R, et al. Machine learning with adversaries: Byzantine tolerant gradient descent[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 118-128.

[38] Yin D, Chen Y, Kannan R, et al. Byzantine-robust distributed learning: Towards optimal statistical rates[C]//International Conference on Machine Learning. PMLR, 2018: 5650-5659.

[39] Pillutla K, Kakade S M, Harchaoui Z. Robust aggregation for federated learning[J]. arXiv preprint arXiv:1912.13445, 2019.

[40] Chen Y, Su L, Xu J. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent[J]. Proceedings of the ACM on Measurement and Analysis of Computing Systems, 2017, 1(2): 1-25.

[41] Li L, Xu W, Chen T, et al. Rsa: Byzantine-robust stochastic aggregation methods for distributed learning from heterogeneous datasets[C]//Proceedings of the AAAI

Conference on Artificial Intelligence: volume 33. 2019: 1544-1551.

[42] Alistarh D, Allen-Zhu Z, Li J. Byzantine stochastic gradient descent[J]. arXiv preprint arXiv:1803.08917, 2018.

[43] Cao X, Fang M, Liu J, et al. FLTrust: Byzantine-robust Federated Learning via Trust Bootstrapping[J]. arXiv preprint arXiv:2012.13995, 2020.

[44] Fang M, Cao X, Jia J, et al. Local model poisoning attacks to byzantine-robust federated learning[C]//29th {USENIX} Security Symposium ({USENIX} Security 20). 2020: 1605-1622.

[45] Chen L, Wang H, Charles Z, et al. Draco: Byzantine-resilient distributed training via redundant gradients[C]//International Conference on Machine Learning. PMLR, 2018: 903-912.

[46] Rajput S, Wang H, Charles Z, et al. DETOX: A redundancy-based framework for faster and more robust gradient aggregation[J]. arXiv preprint arXiv:1907.12205, 2019.

[47] Data D, Song L, Diggavi S N. Data encoding for byzantine-resilient distributed optimization[J]. IEEE Transactions on Information Theory, 2020, 67(2): 1117-1140.

[48] Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images[R]. Technical report, University of Toronto, 2009.

[49] He K , Zhang X , Ren S , et al. Deep Residual Learning for Image Recognition[J]. IEEE, 2016.

攻读研究生期间的研究成果

1. Cao D , Chang S , Lin Z , et al. Understanding Distributed Poisoning Attack in Federated Learning[C]// 2019 IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS). IEEE, 2019.

2. 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(7).

致谢

两年半的研究生生活转瞬即逝，特别感谢我的导师常姗老师在学习和科研上的耐心教导。还记得刚入学时我在科研上十分迷茫，是每周组会上一篇篇顶级会议论文的阅读和学习让我对研究方向慢慢有了想法。常姗老师悉心的教诲使我受益良多，在我完成毕业论文的每个阶段，常姗老师在各个环节中都给予了我耐心的指导。从第一个学期开始，常姗老师就培养我们阅读论文和做学术研究的习惯，在生活、科研方面给予了我耐心的照顾和关怀，每次与常姗老师的交流都能使我受益匪浅，让我学习到了做学术研究应该拥有的严谨态度和研究热情。

感谢东华大学为我们提供了舒适的学习环境和浓厚的学习氛围，能让我在这里专心做好自己的科研论文。也感谢计算机学院的老师们在各个阶段对我的论文提出的改进建议，让我能够更好地地完成论文。

同时也要感谢实验室的小伙伴们，在完成毕业论文的过程中，是他们在方方面面给予我帮助。在遇到科研上的问题时，我们会一起交流探讨，共同碰撞思想的火花。当学习感到迷茫时，是他们帮助我寻找新的方向和动力。生活中，我们也会一起分享快乐和喜悦。

还要感谢我的父母和朋友们，是他们让我在研究生阶段保持积极向上的心态完成学业，专注学习。

在未来的日子里，我会不忘初心，更加努力地过好每一天，不负韶华。