兰州理工大学

硕士论文

学校代号 10731　　学　　号 192085211013

分 类 号 TP309.2　　密　　级 公　开

蘭州理工大學

LANZHOU UNIVERSITY OF TECHNOLOGY

全日制工程硕士学位论文

基于联邦学习的后门攻击研究

学位申请人姓名 李凤强

培　养　单　位 计算机与通信学院

导师姓名及职称 曹来成 教授

学　科　专　业 计算机技术

研　究　方　向 网络与信息安全

论文提交日期 2022 年 03 月 16 日

学校代号：10731
学　　号：192085211013
密　　级：公　开

兰州理工大学全日制工程硕士学位论文

基于联邦学习的后门攻击研究

学位申请人姓名：　李凤强
导师姓名及职称：　曹来成　教授
培　养　单　位：　计算机与通信学院
专　业　名　称：　计算机技术
论文提交日期：　2022 年 03 月 16 日
论文答辩日期：　2022 年 05 月 26 日
答辩委员会主席：　沈玉琳　研究员

Research on Backdoor Attack in Federated Learning

by

LI Fengqiang

B.E. (Shaanxi University of Science & Technology) 2006

A thesis submitted in partial satisfaction of the

Requirements for the degree of

Master of Engineering

in

Computer Technology

in the

School of Computer and Communication

Lanzhou University of Technology

Supervisor

Professor CAO Laicheng

May, 2022

兰州理工大学

学位论文原创性声明

本人郑重声明：所呈交的论文是本人在导师的指导下独立进行研究所取得的研究成果。除了文中特别加以标注引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写的成果作品。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律后果由本人承担。

作者签名：李凤强　　日期：2022 年 5 月 10 日

学位论文版权使用授权书

本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，允许论文被查阅和借阅。本人授权兰州理工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。

本学位论文属于

1、保密□，在_____年解密后适用本授权书。

2、不保密☑。

（请在以上相应方框内打"√"）

作者签名：李凤强　　日期：2022 年 5 月 10 日

导师签名：[illegible]　　日期：2022 年 5 月 10 日

目　录

摘　　要

联邦学习可以让数据在不出本地的情况下参与模型训练，从而得到更好的模型并保护了数据和隐私安全。在《数据安全法》等旨在保护数据和隐私安全的法律法规不断出台的大背景下，联邦学习日益受到重视和广泛应用。但是联邦学习在应用过程中也容易遭到攻击，特别是后门攻击。研究联邦学习中的后门攻击方式，可以促进对于联邦学习自身安全的重视，促进联邦学习更加安全的发展。

本文首先提出一种高效的正反向分裂迭代算法。本文提出的基于特征的联邦学习后门攻击主要是通过正反向分裂迭代算法使得生成的攻击图片在特征上趋向于目标图片，而在外观上基本保持不变。原有的正反向分裂迭代算法使用的是基于范数的损失函数，本文提出了一种新的基于向量的损失函数，在实验比较中发现基于向量的损失函数同等条件下效果优于基于范数的损失函数。

本文其次提出了基于特征的联邦学习后门攻击方案。使用 CIFAR-10 数据集和 ResNet18 模型针对以恶意参与者参与正常训练的数据作为后门和训练过程中植入后门两种不同场景进行了研究。本文提出的基于特征的联邦学习后门攻击方案主要是通过正反向分裂迭代算法使得生成的攻击图片在特征上趋向于目标图片，而在外观上基本保持不变。攻击图片在通过模型识别时由于特征上趋向于目标图片因此容易被分类为目标图片的类别，由于外观上与原始图片基本一致，这样在人工检测等方式下不容易被发现，保证了攻击的隐蔽性。特别在以恶意参与者参与正常训练的数据作为后门的场景下，恶意参与者只需要正常参与训练，不需要做恶意行为，不容易被检测机制判断为恶意参与者，从而提高了恶意参与者参与联邦学习的概率，从而增加了攻击的成功率和隐蔽性。基于特征的联邦学习后门攻击方案在上述提到的两种不同场景下分别讨论了具体的实现方式。在使用目标数据时，提出了目标特征值平均法和目标数据相近法两种方法，通过实验对两种不同的方法的实验结果进行了比较。

本文最后提出了基于生成对抗网络和特征的联邦学习后门攻击方案。本文研究了如何在联邦学习中利用生成对抗网络生成指定标签的数据，并以生成的数据为目标数据，利用基于特征的攻击方法生成攻击数据，从而产生一种新的利用生成对抗网络生成攻击数据的后门攻击方法。这种攻击方式更加灵活和隐蔽。恶意参与者只需要正常参与联邦学习，不需要在训练中做任何恶意行为，提升了参与训练的概率，从而提高了攻击成功率。

关键词：联邦学习；后门攻击；特征提取；隐私保护；生成对抗网络。

Abstract

Federated learning enables participants to construct a better model without sharing their private local data with each other. In the context of the continuous introduction of laws and regulations aimed at protecting data and privacy security, such as the "Data Security Law", federated learning has been more valued and more widely used. However, federated learning is vulnerable to attacks, one of which is backdoor attack. Research on attack methods can promote the attention to the security of federated learning and promote the development of secure federated learning.

Firstly, this thesis proposes an efficient forward-backward-splitting iterative procedure. The federated learning backdoor attack scheme based on feature which proposed in this thesis mainly uses the forward backward splitting iterative procedure to generate the attack image which tend to the target image in feature, but remain basically unchanged in appearance. The original forward backward splitting iterative procedure uses the loss function based on norm. This thesis proposes a new loss function based on vector. In the experiment, it is found that the effect of loss function based on vector is better than that of loss function based norm under the same conditions.

Secondly, this thesis proposes a backdoor attack scheme based on feature. Using cifar-10 data set and resnet18 model, this thesis researches two different scenarios: taking the data of malicious participant participating in normal training as backdoor, another is implanting backdoor in the training. The federated learning backdoor attack based on feature proposed in this thesis mainly uses the forward-backward-splitting iterative procedure to generate attack image based on the feature of target image. The attack images tend to the target image in feature, and basically remain unchanged in appearance. When the attack images are recognized through the model, the attack images tend to the target image in feature, so it will be easily classified as the category of the target image. The attack images remain basically unchanged to the original image in appearance, so it is not easy to be found by manual detection, which ensures the concealment of the attack. Especially in the scenario where the data of malicious participants participating in normal training are used as the back door, malicious participants only need to participate in normal training, and are not easy to be judged as malicious participants by the detection mechanism, which improves the

probability of malicious participants participating in federated learning, increasing the success rate and concealment of the attack. The federated learning backdoor attack based on feature is discussed in the two different scenarios mentioned above. When using target data, two methods are proposed, one is target feature average method and another is target data similarity method. The experimental results of the two different methods were compared.

Lastly, this thesis proposes the federated learning backdoor attack method based on GAN and feature. This thesis discusses how to use GAN to generate the data with specified label in federated learning ,and how to use the attack method based on feature to generate the attack data with take the generated data as the target data, so as to produce a new backdoor attack method using GAN network to generate the attack data. This attack method is more flexible and hidden. Malicious participants only need to participate in federated learning normally and do not need to do any malicious operations in training, which enhances the success rate of participating in training and improves the success rate of attack.

Key words: Federated learning; backdoor attack; Feature extraction; Privacy protection; GAN.

插图索引

附表索引

第1章 绪 论

1.1 研究背景与意义

在大数据时代，数据以海量的方式产生和使用，在传统的数据中心化的方式中需要统一收集数据，然后把数据统一存储到数据服务器或者数据中心、数据湖等地，然后在数据完全可控可见的情况下集中的进行模型训练。但是在有些情形下，想要统一收集数据并不可行。例如在大数据时代，各行各业对数据的重要性都有了非常深刻的认识，同时也都把数据作为公司最重要的核心资产进行保护，不同的公司之间，即使同一公司的不同部门、不同地区之间也会因为各种利益关系和其他原因而不愿意共享自己的数据，因此 “数据孤岛”问题就会日益严重，从而阻碍大数据、人工智能技术的发展，阻碍社会进步。

同时数据中心化的方式还存在着严重的数据泄露、隐私泄露风险。近年来数据信息泄露事件层出不穷，中国电信超2亿条用户数据被贩卖，微博5.38亿用户数据在暗网出售，青岛胶州中心医院6千余人就诊名单泄露，广西医护人员倒卖8万条婴儿信息[1]等等这样的新闻屡见不鲜。同时每个人都有被推销电话不断骚扰的经历，生活中你在一个地方看房留了信息，然后你就会不断的被各种推销房子、推销装修的电话所骚扰，不胜其烦。2021年西宁市公安局重拳出击破获“3.05”侵犯公民个人信息案[2]，此案中被贩卖的公民个人信息达到77万余条。为了有序规范的利用数据，保护数据和个人隐私不被泄露和非法使用，世界各国相继出台了一系列的法律法规。欧盟于2018年颁布的《通用数据保护条例》[3]，中国于2017年颁布的《中国网络安全法》[4]，2021年9月1日起施行的《中华人民共和国数据安全法》[5]。数据安全法开宗明义指出本法旨在规范数据处理活动，保障数据安全，促进数据开发利用，保护个人、组织的合法权益，维护国家主权、安全和发展利益。这些法律法规的出台，对数据的采集和利用进行了严格的限制，原来简单粗放式的收集和利用数据的方式已经不再适用。严格的法律法规倒逼企业创造新的利用数据的方式，一种新的既可以保护数据隐私，又可以联合数据训练模型的全新的数据利用方式被强烈需求。

为了应对以上问题，谷歌公司于 2016 年提出了联邦学习[6]。联邦学习的特点就是可以让各参与者在不需要共享自己数据的前提下联合训练更好的模型，从而保护数据隐私和安全，解决数据孤岛问题。杨强教授在《联邦学习》一书中有一个非常形象的比喻就是羊吃草。在数据中心化模式下，各个草场的草都送到养羊场，然后羊吃草。这里的羊就代表模型，草就代表数据。在联邦学习模式下，草场的草不动，羊被赶到各个草场去吃草，这样草不用动，动的是羊。两种方式下，

羊都吃到了草，不同的是草的运动方式。联邦学习模式下，草不用出草场，保证了草的安全性。由于联邦学习的特点，联邦学习几乎成了数据敏感场景下（如医疗记录、银行信息等）进行模型训练的唯一选择。目前联邦学习已被尝试应用于金融保险、医疗健康、推荐系统、安防系统等多个领域，实验证实这些方法够取得与传统机器学习相近甚至更好的结果。

联邦学习的特点可以让各参与者在不分享自己本地数据的情况下联合其他参与者训练出更好的模型，很好的保护了数据隐私，但是联邦学习的特点也容易遭受各种各样的攻击。由于联邦学习中，各参与者利用本地数据训练模型，服务器端的协调者无法看到参与者的数据和训练过程，就为恶意参与者向数据中投毒和模型中投毒提供了便利条件。在参数聚合过程中的安全又依赖于服务器的安全性，如果服务器遭到攻击，就会影响正常的训练过程或者发生数据泄露。本文研究联邦学习的后门攻击，目的就是为了进一步研究联邦学习中存在的攻击方式和安全漏洞，通过研究攻击方式从而促进联邦学习安全的研究，从而构建更加安全的联邦学习，形成更加安全的联邦学习方案。

1.2 国内外研究现状

联邦学习的特点可以让各参与者在不分享自己本地数据的情况下联合其他参与者训练出更好的模型，很好的保护了数据隐私，但是联邦学习的特点也容易遭受攻击[7-10]。

1.2.1 攻击方式

（1）重构攻击

重构攻击的目的就是使用一定的方法还原出全部或部分数据。文献[11]根据人脸识别模型的训练结果较为准确地还原了原始数据，虽然此文献的研究不是基于联邦学习模式，但文献[12-13]认为该方法同样适用于联邦学习模式。

（2）推理攻击

推理攻击通过推理的方式得出数据中的某一项具体信息。推理攻击分为成员推理攻击和属性推理攻击。成员推理攻击主要有针对性的判断特定数据是否参与训练。属性推理攻击主要有针对性的判断数据集中特定属性是否参与训练。Melis[13]等对FoutSquare数据集上所训练的性别分类器，以 99%的准确率与 100%的召回率实现成员推理攻击。

（3）投毒攻击

文献[14]系统的介绍了在数据中心模式下各种类型的投毒攻击。联邦学习中的恶意参与者由于对本地数据拥有控制权可以通过修改、删除、增加训练数据，以达到破坏原始数据的初始分布和改变学习算法逻辑的目的。主要有两种常见的投

毒攻击的示例：标签翻转攻击[15]和后门攻击[16-21]。

a.标签翻转攻击

标签翻转攻击中恶意参与者可以翻转样本标签，使训练得到的模型偏离既定的预测边界[13]。在联邦学习中参与者利用本地数据训练模型，由于对本地数据拥有绝对的控制权，对本地数据做出任何的修改、删除、增加等操作都是可能的，也不容易被其他用户所感知，这就为恶意参与者进行数据投毒、模型投毒提供了便利条件。甚至有些恶意参与者可以不通过训练数据而是利用算法自适应地生成攻击模型参数或者梯度信息，最大限度地提高投毒攻击的效率。

b.后门攻击

后门攻击需要设计触发器来触发攻击，触发器未触发时不影响正常任务的性能，只有触发器触发时才按照植入后门的要求进行分类[16]。文献[17]指出对于联邦学习框架,攻击者可以利用后门数据对本地模型进行训练,从而在生成的模型中嵌入触发器，当触发器被触发时按照设定好的方式运作产生结果。由于联邦学习常用平均聚合的方式对各参与者提交的模型进行聚合，通过聚合之后会减弱恶意参与者所提交模型的影响，从而导致后门攻击力度减弱。恶意攻击者为了提高攻击率会提交按比例放大的训练结果，这样恶意参与者的模型的在全局模型中的比重增大，甚至可以完全抵消其他参与者的对模型的影响，以至于恶意参与者的模型就是全局模型，从而大大增强后门攻击的力度。后门攻击的特点就是模型对于主要任务的识别率不受影响，可以正常工作，只有当后门触发器触发时才按照设定好的后门工作。

联邦学习后门攻击分为两种类型，一种是语义后门攻击，一种是像素后门攻击[22-23]。在语义后门攻击中攻击者不用修改输入数据，仅是给具有特定特征的数据指定标签,经过训练后具有特定特征的数据就会被识别为指定标签类型的数据。像素后门攻击对输入数据进行修改设置触发器，并指定类别，经过训练后模型会把带有触发器的数据识别为指定类别。比如在CIFAR-10 数据集中，给图片添加红色条块，并指定类别为鸟类，经过训练后，带有红色条块的数据会被识别为鸟类。但是这种构造的触发器容易被人工检测发现，隐蔽性不够好。Xie等人[19]在现有的攻击方式上，提出了分布式后门攻击(distributed backdoor attack ,DBA)，该方案采用多个触发器植入后门，多个局部触发器组合成全局触发器，提升了后门攻击的精度。文献[20-21]两篇文献都提出了一种利用GAN网络生成触发器植入后门的方式，在GAN网络中利用联邦学习的全局模型参数来更新判别器D的参数，由于全局模型参数是由各参与者提交训的练模型聚合而来，相当于在全局数据上来训练判别器。通过生成器G和判别器D来生成和真实数据同分布的假数据，然后恶意参与者把假数据以水印的方式植入到自己的训练数据中，从而达到植入后门的目的，实现后门攻击。由于后门触发器接近于真实数据，所以在植入过程中不容易

被检测，隐蔽性更高，同时收敛速度也会加快。

（4）服务器端 GAN 攻击

Wang[24]等介绍了一中在基于服务器端GAN攻击的攻击方法mGAN-AI，可以获取指定用户的数据。在联邦学习中的服务器引入GAN攻击，一般是在服务器端建立生成对抗网络，用聚合的全局模型来更新GAN网络中的识别器D，然后训练生成器G，从而产生不同类别的数据。mGAN-AI进一步提出一种多任务的GAN网络，识别器要识别真假数据，类别以及参与者身份，从而重构指定参与者的数据，让攻击达到客户级别。

1.2.2 联邦学习隐私保护对策

（1）异常检测与对抗训练

Shen[25]等提出了一种间接协作的深度学习框架，在此框架中各参与者向服务器提交的不是模型梯度而是提交经过精心设计的伪装特征，这种伪装特征可以利用自动统计机制来抵御投毒攻击。在此框架下，协调者使用 K-means 算法对每个参与者提交的伪装特征进行聚类从而检测出异常点。这些检测到的异常点可以进一步用于与攻击策略相关联的伪装特征对恶意用户进行识别[26]。

对抗性训练是一种主动防御技术，就是在训练阶段就把攻击者可能出现的攻击加入训练，这样使得模型对于已经训练过的攻击方式可以很好的抵御，从而对抗已知的攻击。文献[16]讨论了如何通过对抗训练的机器学习模型对攻击活动具有较强的鲁棒性。

（2）模型压缩

模型压缩是深度学习领域常见的一种技巧,主要用于减少模型的参数和大小，提高模型的训练和推断速度。模型压缩导致模型传输的不是原始的参数数据，可以与差分隐私一样，即使恶意攻击者窃取了中间的模型参数，也很难将其还原。

（3）差分隐私

差分隐私在隐私保护方面占有重要的地位和作用。差分隐私主要是通过给敏感隐私数据添加噪声，通过添加噪声干扰的方式从而达到用户数据隐私的保护。而且通过差分隐私大大提高了数据的隐私性，同时对于数据质量不会有太大的干扰和影响，所以差分隐私保护现在成为了隐私保护的主要技术之一。在联邦学习中，为了避免服务器从参与者上传的梯度信息或者模型参数信息推理出更多的数据隐私，利用差分隐私向参与者上传的参数信息中添加噪声，使得服务器无法区分。Abadi等[27]提出了一种将差分隐私保护机制与SGD算法相结合的隐私保护深度学习方法，该方法主要是通过在小批量步骤后利用噪声干扰本地梯度实现隐私保护。差分隐私可以增加数据的安全性，提升隐私保护的能力，同时也会增加开销，降低联邦学习的学习效率，需要在隐私保护和效率之间做出平衡。

（4）安全多方计算

安全多方计算[28-30]（Secure multi-party computation，SMC）目的是为了保护多个参与者在联合训练模型或者计算函数时的输入数据是安全的，不会被其他参与者所获取。在联邦学习中各参与者提交的信息是自己的梯度信息或者是模型参数信息，所以只需要通过安全多方计算来保护这些梯度信息或者模型参数信息，而不是原始数据，大大减小了计算量，提高了效率。这种性能特点使得安全多方计算成为联邦学习环境下的首选技术。Aono等[29]指出，在联邦学习中，恶意参与者仅通过各参与者提交的梯度信息就可以还原出一些额外的信息，使得各参与者的数据安全遭到威胁。为了保护各参与者提交的梯度信息的安全，引入了同态加密技术，使用同态加密技术对参与者提交的梯度等信息进行加密，服务器端的协调者在接收到所有的信息后在同态加密方式下进行聚合，聚合完成后把聚合的全局参数分发各参与方，各参与方解密后更新本地模型，继续进行本地训练。同态加密技术可以保护信息安全，但同时增加了计算开销，在运算量很大的模型中，这种开销的增大是无法忽视的。针对这种情况，Bonawitz等提出了另外一种实用的安全聚合协议[31]，在这个协议中，一个用户在他的私有向量上添加掩码，另一个用户可以使用相同的种子去剔除这些掩码。通过这种方式，服务器可以通过对接收到的梯度向量执行求和操作来消除所有掩码。此外，该协议还考虑了用户退出场景，设计了一个TSS方法来解密平均更新 [32-33]。

（5）同态加密

同态加密的特点是密文运算的结果解密后和明文计算的结果相同，可以利用同态加密技术对联邦学习中参与者和服务器之间交换的信息进行保护，防止恶意服务器重构和泄露信息，保证交换数据的安全。文献[34]提出了一种新的深度学习系统，此系统基于诚实并好奇的云服务器，使用同态加密方案保护提交到云服务器上的梯度，保证了提交的梯度信息的安全，同时保证系统训练精度不降低。

1.3 论文的研究内容

(1)研究高效的正反向分裂迭代法

本文提出的基于特征的联邦学习后门攻击主要是通过正反向分裂迭代算法使得生成的攻击图片在特征上趋向于目标图片，而在外观上基本保持不变。攻击图片在通过模型识别时，由于特征上趋向于目标图片，容易被识别为目标图片的类别标签，由于在外观上基本保持不变，这样在人工检测等方式下不容易被发现，保证了攻击的隐蔽性。在文献[35]中，实现让生成的攻击图片趋向于目标图片特征的算法中，损失函数使用的是范数。本文提出一种新的使用向量的损失函数，在实验中比较发现使用向量的损失函数同等条件下效果优于使用范数的损失函数。

(2)研究联邦学习模式下基于不同场景的后门攻击方法

联邦学习的特点是参与者利用本地数据通过服务器下发的全局模型训练本地模型，模型训练结束后向服务器提交训练好的模型参数或者梯度信息。参与者之间，参与者和服务器之间都是不进行原始数据的交流，这样保证了数据的安全。这种方式下，参与者可以控制训练的过程也可以得到全局训练模型。本文根据这个特点基于两种不同的场景提出了一种新的后门攻击方法。两种场景分别如下，一种是以恶意参与者正常参与训练的数据作为后门。恶意参与者参与联邦学习的过程中拥有自己的本地数据，如果恶意参与者拥有想要攻击的目标类型的数据，就可以以自己拥有的数据作为后门，通过本文提出的基于特征的攻击方法生成攻击图片。以CIFAR-10数据集为例，恶意参与者拥有部分标签为2的鸟类数据，同时恶意参与者希望将任意的测试数据分类为标签为2的鸟类数据。这时候恶意参与者在训练过程中不用植入后门只需正常的参与训练，等到模型训练完成后，把需要分类的数据通过本文提出的方法让其在特征上趋向于恶意参与者拥有的鸟类数据特征，而在外观上保持基本不变。这种方法可以通俗的理解成就是把恶意参与者拥有的数据作为后门来使用，这种方式不用改变恶意参与者的训练过程，隐蔽性极高。第二种场景就是恶意参与者不拥有需要攻击的目标数据，需要植入后门数据。这种情形下，可以通过植入部分其他数据，标签设置为目标标签进行训练。需要攻击模型时，以植入的数据为目标来处理待分类数据从而生成攻击数据。

(3)研究基于生成对抗网络 GAN 和特征的后门攻击方法

生成对抗网络就是一个正方和敌方相互博弈的过程，相互促进，共同成长最后达到一个平衡。具体就是有一个生成器G和一个判别器D，G负责生成更加逼真的假数据来欺骗D，D的职责是不断训练提高判断假数据的能力从而可以非常准确的区分假数据和真数据，最终产生的结果是G生成的假数据D无法判断真假，只能猜测，真假概率分别为50%。现有的研究是在联邦学习中通过生成对抗网络GAN来生成触发器，然后把触发器植入模型从而达到后门攻击的目的。本文研究把GAN网络生成的生成数据和基于特征的攻击方法相结合，从而产生一种新的利用GAN网络生成攻击数据的后门攻击方法。

1.4 论文的组织结构

本文主要有以下 5 个章节：

第 1 章 绪论。本章简要的描述了全文的组织结构与研究内容。并对课题的研究背景、研究意义以及国内外研究现状进行的详尽的叙述。

第 2 章 相关理论与技术。本章详细介绍了联邦学习的基本概念和分类、联邦学习中的后门攻击方式、联邦差分隐私、同态加密技术、ResNet18 模型、模型特征提取的方法、GAN 网络的基本知识、Pytorch 中梯度下降的实现方法。

第 3 章 基于特征的联邦学习后门攻击。介绍了生成攻击数据的正反向分裂迭

代算法，列出了算法的两种具体实现过程，比较了两种算法的差异。详细说明了在两种不同场景下如何使用目标数据，如何在选定使用目标数据的方法后使用算法生成攻击数据，从而实现对模型的攻击。最后介绍了具体的实验过程和实验结果。

第 4 章 基于生成对抗网络 GAN 和特征的联邦学习后门攻击方法。介绍了恶意参与者如何利用生成对抗网络生成具有指定标签的假数据，以及如何利用这些假数据在基于特征的攻击方法下生成最终的攻击数据。最后介绍了具体的实验过程和实验结果

最后总结与展望。总结本文所做的各种实验结果与分析论断，并展望一下尚待完善的部分与未来的重点研究方向。

1.5 本章小结

本章主要介绍了本文研究的背景与意义，主要是介绍了联邦学习在隐私保护中的重要意义和建立安全的联邦学习的重要性。介绍了国内外研究现状，在现状中主要讲了两部分内容，一部分是联邦学习中的攻击方式，一部分是联邦学习中的保护对策，一矛一盾让大家对于联邦学习自身的安全有一个直观的了解。介绍了论文的研究内容，包括生成攻击数据的算法，基于特征的后门攻击方法，基于生成对抗网络和特征的后门攻击方法。

第2章 相关理论介绍

2.1 联邦学习

2016 年谷歌为了解决安卓手机终端用户在本地更新模型的问题，提出了一种被称为联邦学习的设计方案，联邦学习的目标是保护数据交换时的安全，保护数据隐私，保证数据在遵守法律法规的前提下被合法使用，在满足这些条件的基础上在多个参与方之间联合训练新的模型。联邦学习经常被用于神经网络，但实际上联邦学习不局限于神经网络，还可以应用于随机森林等算法。联邦学习更像是一种新的安全平台，可以在其上实现各种各样的人工智能算法，联邦学习有望成为下一代人工智能协同算法和协作网络的基础。

联邦学习的特点就是可以让各参与者在不需要共享自己数据的前提下联合训练更好的模型，从而保护数据隐私和安全，解决数据孤岛问题。杨强教授在书中有一个非常形象的比方就是羊吃草。在训练数据中心化模式下，各个草场的草都送到养羊场，然后羊吃草。这里的羊就代表模型，草就代表数据。在联邦学习模式下，草场的草不动，羊被赶到各个草场去吃草，这样草不用动，动的是羊。两种方式下，羊都吃到了草，不同的是草的运动方式，联邦学习模式下，草不用出草场，保证了草的安全性。

设矩阵 Di 表示第 i 个参与方的数据，定义矩阵 Di 的每一行表示一个数据样本，每一列表示一个具体的数据特征。我们将数据的特征空间记为 F，数据标签空间记为 L，数据样本的 ID 空间记为 I，这样就组成了一个训练数据集（I,F,L）。根据训练数据在不同参与方之间的数据特征空间和样本 ID 空间的分布情况，我们将联邦学习分为三种类型，分别介绍如下。

2.1.1 横向联邦学习

首先举个例子直观的来了解横向联邦学习的特点。比如，A 市和 B 市有两家做同类保险业务的保险公司，由于业务相近，客户的数据可能会拥有非常相似的特征，但是由于分属不同的市，用户可能只会有少量的重叠。这意味这两家保险公司用户的重叠部分较小，数据特征部分重叠较大，两家公司可以通过横向联邦学习来协同建立一个学习模型，模型如图 2.1 所示。

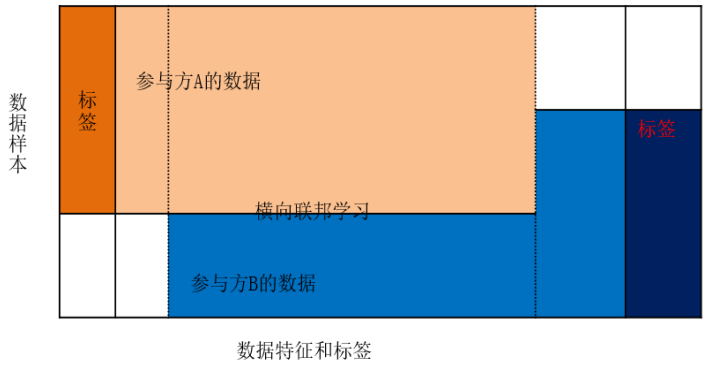


图 2.1 横向联邦学习样本划分

横向联邦学习系统的典型模式就是客户-服务器架构，如图 2.2 所示。在这种模式下，K 个拥有同样数据特征的参与者联合一起训练模型，由服务器中间协调 K 个参与者协同工作，具体过程包含下面几步：

步骤 1.各参与方接收服务器发来的全局模型，然后在本地利用自己的本地数据训练模型，可以使用同态加密、差分隐私或秘密共享等加密技术来保护模型参数或者梯度信息，并将受保护的梯度信息或者模型参数信息发送给聚合服务器。

步骤 2 服务器接收各参与者发送的更新模型参数或者梯度信息，然后对其进行安全聚合，安全聚合包括基于同态加密的加权平均[36-37]等方法。

步骤 3 服务器将聚合后的结果发送给各参与方。

步骤 4 各参与方接收服务器发送的梯度信息或者模型参数信息，并使用这些信息更新模型参数。

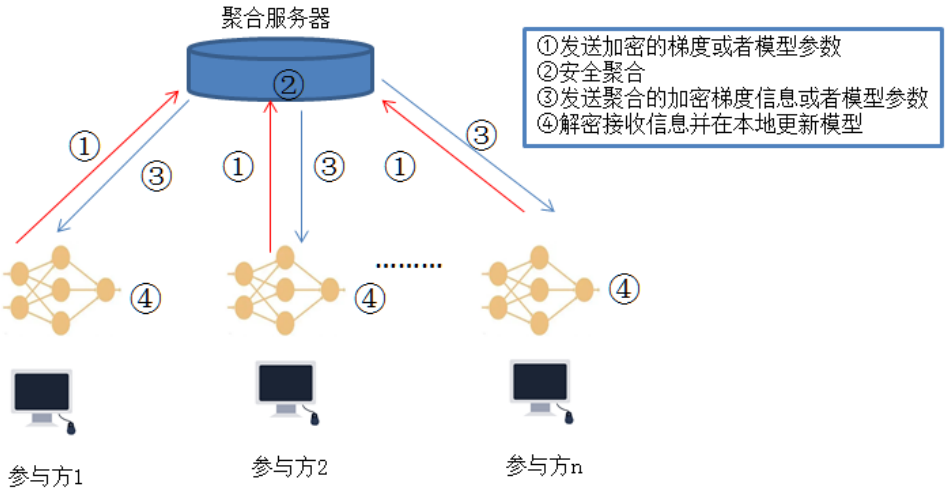


图 2.2 横向联邦学习训练过程

在上述步骤中聚合服务器有两种聚合方式，一种称为梯度平均[38-40]，一种称为模型平均[38,41]。梯度平均的过程是这样的，各参与方在完成本地模型训练后，将得到的梯度信息发送给服务器，服务器在收到梯度信息后进行聚合，然后再将聚合后的梯度信息发送给参与方，各参与方接收到全局梯度信息后更新本地

模型，以新模型继续本地训练。模型平均的过程是这样的，各参与方完成本地训练后，将模型的参数信息发送给服务器，服务器在收到模型参数信息后进行聚合，然后再将聚合后的模型参数信息发送给参与方，各参与方接收到全局模型参数信息后更新本地模型，以新模型继续本地训练。

下面我们将介绍最关键的联邦平均算法。

算法 2.1 联邦平均算法
在协调方执行： 初始化模型参数 W_0，给所有参与方发送模型参数 W_0 for i = 1 to N:#服务器端迭代次数为 N 协调方在参与方中随机选取 K 个参与方形成参与方集合 C_t for c in C_t:#对 C_t 中所有的参与方进行处理 参与方通过本地训练更新本地模型 w_{i+1}^c。 将更新后的模型参数 w_{i+1}^c 发给协调方 end for 协调方将收到的模型参数，通过加权平均的方式进行聚合： $\tilde{W}_{i+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} W_{i+1}^k$ 协调方检查是否收敛，如果收敛则停止训练。 没有收敛协调方将聚合后的模型参数 $\tilde{W}_{i+1}$
在参与方更新 参与方属于 C_t 集合，ID 为 k，轮数为 i 接收服务器广播的全局模型参数，假设 $W_i^k = \tilde{W}_i$ for j =1 to E:#本地迭代 E 轮 本地训练模型更新参数 W_K^{i+1} end for 将训练完成的本地模型参数 W_K^{i+1} 发送给协调方进行聚合。

2.1.2 纵向联邦学习

首先举个例子直观的来了解纵向联邦学习的特点。比如，A 市有两家公司，一家是银行，一家是保险公司，两家公司在客户群体上有很大的交集，但是业务特征信息有很大不同。这意味这两家公司用户的重叠部分较大，数据特征部分重叠较小,两家公司可以通过纵向联邦学习来协同建立一个学习模型,如图 2.3 所示。

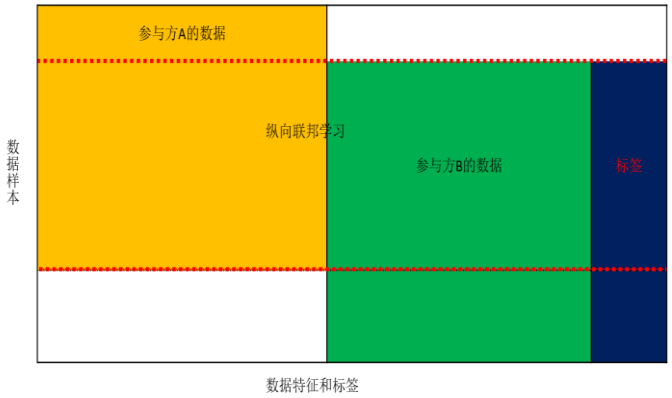


图 2.3 纵向联邦学习样本划分

纵向联邦学习在训练过程中一般包括两个大的步骤，首先对参与方中拥有相同 ID 的数据进行对齐，其次对这些对齐的数据进行联邦学习。

步骤 1 加密实体对齐

A 公司和 B 公司通过纵向联邦学习方式进行联合训练模型，由于在不同公司中用户存储位置不同，需要先把用户对齐，这样才能把分布在两个公司的同一用户的特征联合起来训练模型。其中使用的是被称为基于加密的用户 ID 对齐技术，可以参考文献[42-43]所描述的，确保 A 公司和 B 公司在不用共享用户信息的情况下对齐共同用户。在对齐过程中，每个公司的用户信息都是受隐私保护的，不会泄露。

步骤 2 加密模型训练

在对齐共有用户信息后各参与方可以联合训练模型，训练过程分为四步。

第一步协调者 S 创建秘钥对，并将公钥发给 A 公司和 B 公司。

第二步 A 公司和 B 公司加密中间结果并交换，这个中间结果用来帮助计算梯度和损失值。

第三步 A 公司和 B 公司计算加密梯度，然后加入掩码。B 公司同时计算加密损失，A 公司和 B 公司将加密的结果发送给协调者 S。

第四步 S 方对梯度和损失进行解密，并将结果发送给 A 公司和 B 公司。A 公司和 B 公司解除梯度信息上的掩码，并根据这些梯度信息更新模型参数。

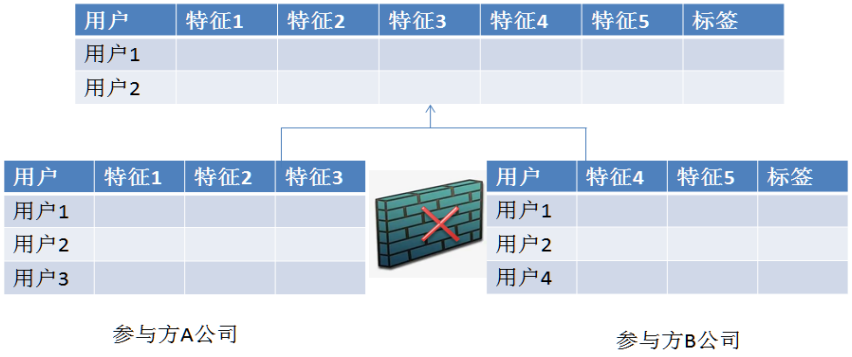


图 2.4 样本对齐

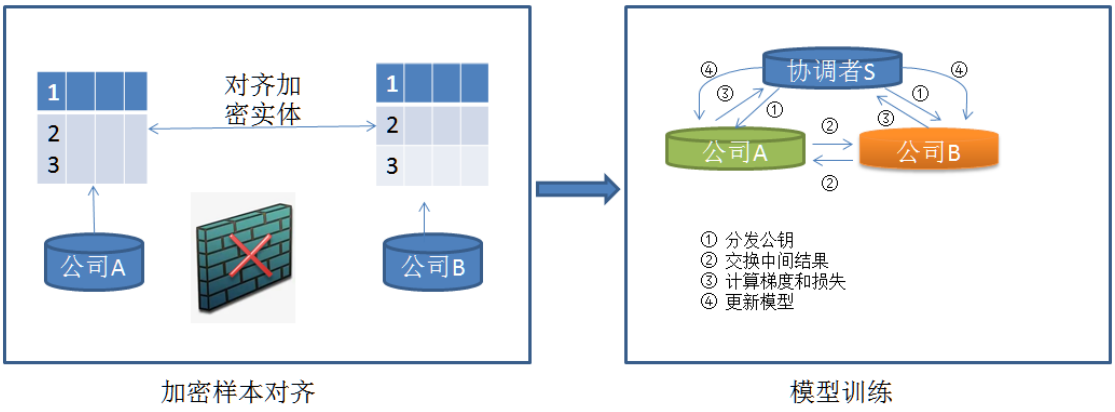


图 2.5 纵向联邦学习训练过程

2.1.3 联邦迁移学习

参与方在用户部分交集很少，在数据特征部分交集也很少，能不能用联邦学习来联合训练更好的模型，答案是肯定的，可以通过联邦迁移学习来协同建立学习模型。

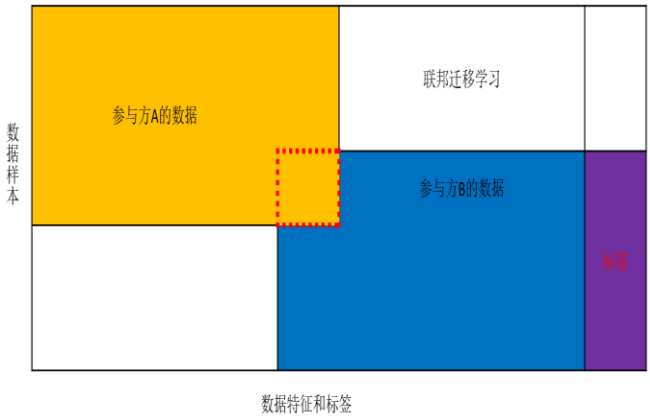


图 2.6 迁移联邦学习样本划分

2.2 后门攻击

在联邦学习中，后门攻击非常普遍，因为后门攻击相对隐蔽，后门攻击不影响模型的主要任务，只有当模型检测到植入的后门特征时才会触发后门触发器，按照后门设定好的方式进行工作。联邦学习中由于有多个参与者共同训练模型，而且每个参与者都在本地利用自己拥有的本地数据训练模型，然后通过参数服务器安全聚合后形成全局模型。这种方式就为恶意参与者攻击模型提供了便利，恶意参与者在本地利用有毒数据训练模型，把后门特征植入本地模型，然后通过全局聚合把后门特征植入全局模型。恶意参与者植入后门算法如算法 2.2 所示。

算法 2.2　恶意参与者植入后门算法

输入:客户端 ID：k,全局模型 θ_G，学习率：η，本地迭代次数:E,
每一轮训练样本：batch_size
输出：返回模型更新的差值：r(P-θ_G)
开始：
用全局模型参数 θ_G 更新本地模型 P;
损失函数:$Loss=Loss_{class}+Loss_{distance}$
$Loss_{class}$ 模型预测和真实结果之间的差值
$Loss_{distance}$ 本地模型和全局模型参数之间的差值
for i =1 to E：
将本地数据切分为大小为 batch_size 的数据集合 B；
for b in B:
用数据块 b 使用梯度下降训练模型:
$P = P - \eta \nabla Loss$
End for
End for
return r(P-θ_G)

恶意参与者训练的目的就是要让模型在正常数据上分类的精度不变，然后在具有后门特征的数据上精度尽可能的高。联邦学习由于会对每个参与者提交的模型进行聚合，聚合的方式以 FedAVG 为主，这种聚合方式以求平均值的方式对每个参与者提交的模型参数进行处理，这样恶意参与者提交的有毒模型就会被稀释掉，后门攻击性能就会下降，为了尽可能的提高恶意参与者的后门攻击能力，就需要尽可能的提高恶意参与者在全局模型中的比重，推导过程如下：

（1）正常的模型聚合过程如下：

$$G^{t+1} = G^t + \frac{\eta}{n}\sum_{i=1}^{m}(P_i^{t+1} - G^t) \qquad (2.1)$$

G^t 代表第 t 轮的全局模型，P_i^{t+1} 代表第 i 个参与者 t+1 轮的本地模型，n 代表参与者的个数。

（2）对于恶意参与者 C_m 期望的就是自己训练得到的模型 P 就是全局模型，P_m^{t+1} 为恶意参与者 C_m 提交的模型，即：

$$P = G^t + \frac{\eta}{n}\sum_{i=1}^{m}(P_i^{t+1} - G^t) \Rightarrow P = G^t + \frac{\eta}{n}\sum_{i=1}^{m-1}(p_i^{t+1} - G^t) + \frac{\eta}{n}(P_m^{t+1} - G^t) \Rightarrow$$

$$P_m^{t+1} = \frac{n}{\eta}P - \sum_{i=1}^{m-1}(p_i^{t+1} - G^t) + (\frac{n}{\eta} - 1)G^t$$

(3)由于当模型接近于收敛时，$\sum_{i=1}^{m-1}(p_i^{t+1}-G^t)$趋向于 0，所以上面的式子可以近似表示为：

$$P_m^{t+1}=\frac{n}{\eta}(P\text{-}G^t)+G^t \qquad (2.2)$$

(4)上面的式子中η小于 n,所以实质就是放大恶意参与者的模型，让在聚合过程中无法抵消恶意参与者模型的作用，从而达到提升攻击效果的目的。

2.3 联邦差分隐私

在联邦学习中，可以利用差分隐私来保护数据安全和用户安全。传统的差分隐私中的相邻数据集定义为：

相邻数据集：设有两个数据集 D 和 D，若它们之间有且仅有一条数据不一样，那我们就称 D 和 D 为相邻数据集。

在联邦学习中那些用户参与了训练也是隐私，不能被其他用户推断得到。所以定义了用户相邻数据集的概念。

用户相邻数据集：设用户 C_i 对应的本地数据集为 d_i，D 和 D 代表两个不同用户集合拥有的数据集合，当且仅当 D 去除或者添加某一个客户端 C_i 的本地数据集 d_i 后变为 D，这时我们称 D 和 D 为用户相邻数据集。

算法 2.3 联邦学习客户端差分隐私算法

输入：客户端 ID：K，全局模型：θ_G，模型参数：学习率η，本地迭代次数 E，每一轮训练样本大小：batch_size。
输出：本地模型更新差值：P-θ_G
模型训练开始
使用全局模型θ_G更新本地模型 P，P=θ_G
for i =1 to E：
将本地数据切分为大小为 batch_size 的数据集合 B；
for b in B:
用数据块 b 使用梯度下降训练模型:$P=P-\eta\nabla Loss$
参数裁剪：P=θ_G+clip(P-θ_G)
End for
End for
return P-θ_G
模型训练结束

算法 2.4　联邦学习服务器端差分隐私算法

输入：学习率 η,梯度裁剪边界值:C,噪声参数： σ

随机初始化全局模型参数 θ_G

定义客户端权重：客户端 C_i 对应的权重 $W_k = \min(\frac{n_k}{w}, 1)$

设 $W = \sum_k W_k$

for i=1 to K:#迭代 K 次

以概率 q 挑选参与训练的本轮训练的客户端集合 C^t

for c in C^t:

执行本地训练：客户端 C 执行联邦学习客户端差分隐私算法 2,得到更新值 Δ_c^i

end for

服务器聚合客户端参数：

$$\Delta^i = \begin{cases} \dfrac{\sum_{c \in C^t} W_c \Delta_c^i}{qW} & \text{for } \tilde{f}_f \\ \dfrac{\sum_{c \in C^t} W_c \Delta_c^i}{\max(qW_{\min}, \sum_{c \in C^t} W_c)} & \text{for } \tilde{f}_c \end{cases}$$

对 Δ^{i+1} 的值进行裁剪： $\Delta^{i+1} = \Delta^{i+1} / \max(1, \frac{\| \Delta^{i+1} \|}{C})$

求取高斯噪声的方差：

$$\sigma = \begin{cases} \dfrac{zS}{qW} & \text{for } \tilde{f}_f \\ \dfrac{2zS}{\max(qW_{\min})} & \text{for } \tilde{f}_c \end{cases}$$

更新全局模型参数： $\theta_i = \theta_{i-1} + \Delta^i + N(0, I\sigma^2)$　# $N(0, I\sigma^2)$ 高斯分布

2.4 同态加密

1978年Rivest[44]等人提出了同态加密的概念。同态加密的特点是就是可以直接在加密后的密文上直接计算，而且计算得到的结果解密后和明文直接进行计算得到的结果是一样的。一个同态加密流程如图2.7所示，首先把明文加密得到密文，然后在密文的基础上进行运算得到密文运算的结果，对结果解密后就可以得到和明文直接计算同样的结果。基于同态加密这种特性，在联邦学习中可以用同态加密方法加密梯度信息或模型参数信息，然后服务器的协调者在密文上进行聚合，这样传递的梯度信息或模型参数信息就得到了保护。

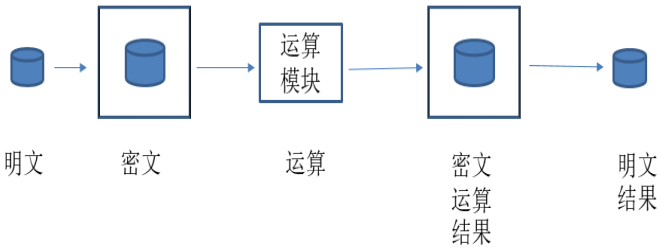


图 2.7 同态加密

2.4.1 同态加密的定义

同态加密方案 HE 包括四部分内容

HE={KeyGen, Enc, Dec, Eval}

KeyGen 表示秘钥生成函数。KeyGen 可以生成非对称的秘钥对{PubKey,PrivateKey},也可以生成对称秘钥key,生产的秘钥用来加密和解密。Enc表示加密函数、Dec表示解密函数、Eval表示评估函数。

加法同态运算：对于在明文空间 M 中的任意两个元素 u 和 v，其加密结果分别为 $E(u)$ 和 $E(v)$,满足：

$$Dec(E(u)+E(v))=Dec(E(u+v))=u+v \quad (2.3)$$

乘法同态运算：对于在明文空间 M 中的任意两个元素 u 和 v，其加密结果分别为 $E(u)$ 和 $E(v)$,满足：

$$Dec(E(u)*E(v))=Dec(E(u*v))=u*v \quad (2.4)$$

2.4.2 同态加密分类:

同态加密方法可以分为三类：部分同态加密，些许同态加密，全同态加密。

（1）部分同态加密

部分同态加密方案是只满足加法同态或者乘法同态，而且操作符可以无限次地用于密文。

（2）些许同态加密

些许同态加密指的是经过同态加密后的密文只能进行有限次的运算。密文上的每一次操作都会增加密文上的噪声，当噪声达到一定的阈值之后，解密操作就无法正常运算，无法得到正确结果，同态加密方案就会失效，所以些许同态加密会限制计算操作次数。

（3）全同态加密

全同态加密算法允许对密文进行无限次地加法和乘法运算操作。任何函数都可以转换为只包含加法和乘法的操作,所以任何函数都可以构造使用全同态加密。当前，全同态加密仍在高速发展，在高效的自举算法[45-47]、多秘钥全同态加密[48]等领域依然是研究热门领域。目前全同态加密建立在些许同态加密方法基础上，

并通过代价高昂的自助法操作实现，但是自助法的代价高昂，因此全同态加密方案计算十分缓慢且在实践中并不比传统的安全多方计算方法更好，因此很多研究人员目前正着眼于发现满足特定需求的些许同态加密方案，而非发掘全同态加密方案。

2.4.3 同态加密方案在联邦学习中的应用

算法 2.5 客户端加入同态加密方案的联邦学习算法
输入：全局模型加密参数 E_{Gp}, 用全局模型权重更新本地模型权重 E_{Lp}<—E_{Gp} for i=1 to E :#E 代表迭代次数 选取 batch_size 大小的训练数据进行训练 在模型参数加密状态下求解加密梯度 利用加密梯度更新模型的加密参数 E_{Lp} end for return E_{Lp}

2.5 ResNet18 模型

在训练卷积神经网络的过程中，如果训练效果较差时人们自然会想到能不能通过加深网络成熟来提升模型性能。理论上随着网络深度的增加，模型可以表达的信息就会更丰富，效果就会更好。但是在实验中科学家发现增加模型的深度刚开始可以降低模型的损失值，但是随着深度的增加，模型的损失值不降反升了，也就是网络性能出现了退化现象。这个现象困扰着科学家，直到何凯明等人提出了一种新的网络结构才解释了网络退化的原因，也才找到了解决问题的办法，这就是 ResNet[49]，ResNet 模型结构如图 2.8 所示。

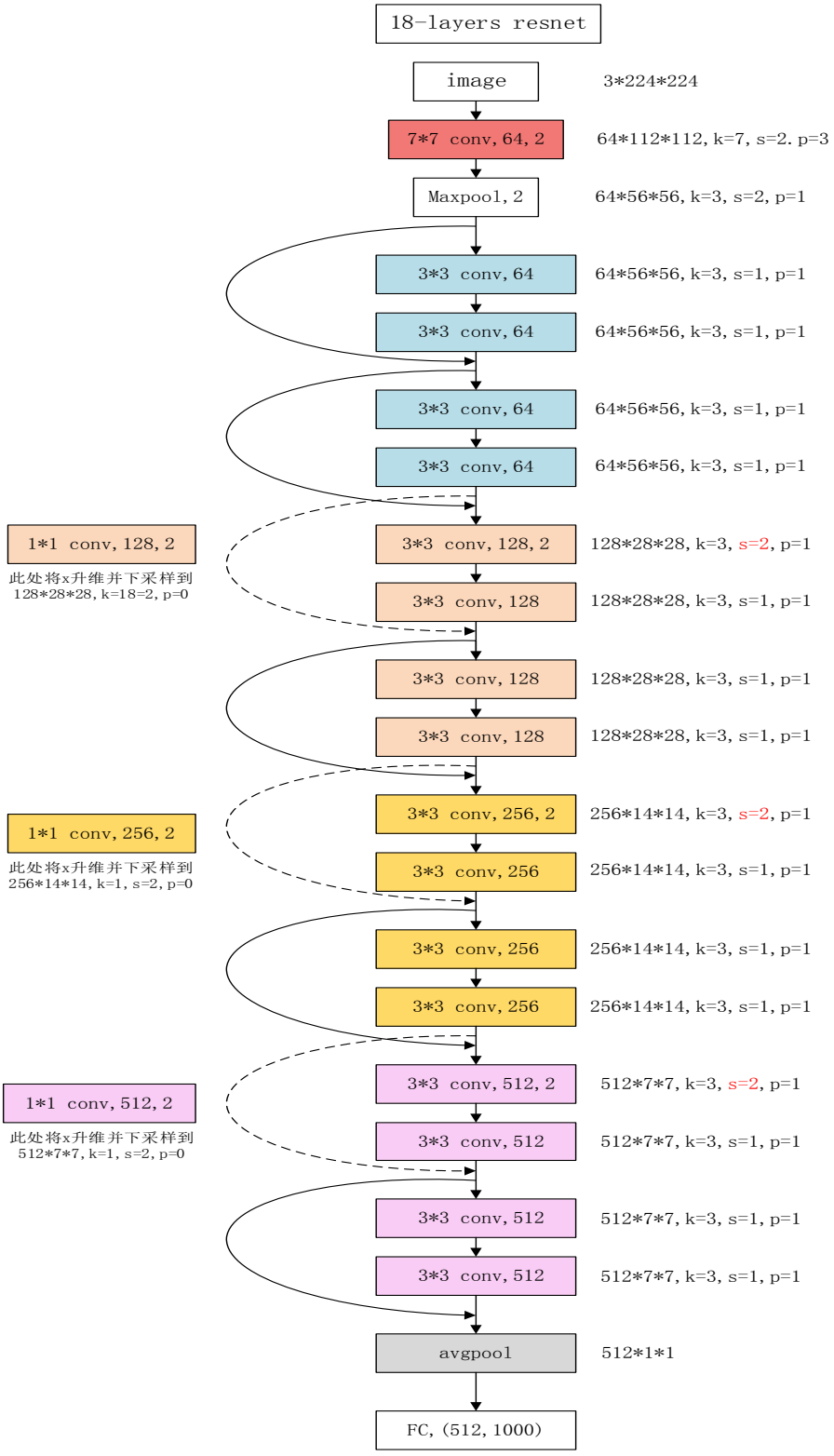


图 2.8 ResNet18 结构

为了进一步理解 ResNet 模型，我们先介绍一下 ResNet 的基本单元:残差块

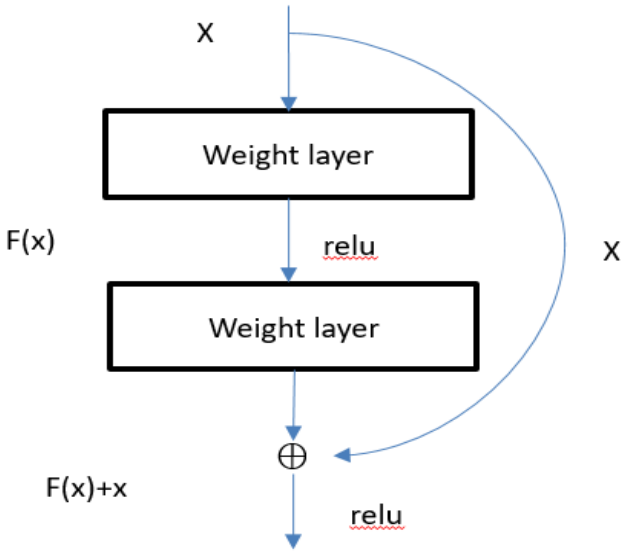


图 2.9 残差块

ResNet 的残差块在一个或者多个卷积层之间加上一条“短接线”（shortcut connection），这条“短接线”会起到一个恒等映射（identity mapping ）的作用。接下来仔细介绍一下，以这种结构组成的 ResNet 网络为什么可以避免深层网络的“退化”现象：假设输入为 X，将以 X 为输入的某网络层的输出设为 H(X)。在一般的卷积神经网络如 VGG 中，由输入 X，需要通过网络层的训练直接拟合得到输出 H(X)；而在上图所示的基本单元中，由于存在短接线，由输入 X，为得到输出 H(X)，只需要通过网络层的训练拟合得到残差函数 F(X)=H(X)-X，再间接得到基本单元的输出 H(X)=F(X)+X。那么，为什么我们要拟合残差函数 F(X)=H(X)-X，而不是直接拟合得到 H(X)呢?现在我们假设网络达到某一深度的时候，该层的输出 X 已经达到最优状态，也就是说，此时的错误率是最低的时候，再继续加深网络的层数就会出现退化的现象。如果使用一般的卷积神经网络，现在将该层的输出 X 作为下一层的输入进行训练，这时更新下一层的权值的代价相对较高，因为下一层的权值要使得输出 H(X)仍然保持最佳状态，即 H(X)=X。但是采用 ResNet 时，还是假设网络达到某一深度的时候，该网络的输出 X 已经达到最优状态，将该层的输出 X 作为下一层的输入进行训练，为了保证下一层的输出 H(X)仍然是最优状态，只需要拟合残差函数 F(x)=H(X)-X=0 就可以了，这时下一层的输出 H(X)就等于 X。当然上面提到的只是理想情况，但是总会有那么一个时刻，某一层的输出能够无限接近最优解。所以，相对于使用一般的卷积神经网络如 VGG，需要直接拟合得到输出 H(X)，而采用 ResNet 只需要拟合残差函数 F(x)，而得到残差函数 F(x)只需要更新 F(x)少部分的权值就可以了。

ResNet18 模型参数如图所示：

Layer name	Output size	18-layer
Conv1	112×112	7×7,64, stride 2
Conv2_x	56×56	3×3, max pool, stride 2
		$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times2$
Conv3_x	28×28	$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times2$
Conv4_x	14×14	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times2$
Conv5_x	7×7	$\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times2$
	1×1	average pool, 1000-d fc, softmax

图 2.10 ResNet 模型参数

2.6 模型提取数据特征

神经网络的结构是一层一层的，上一层的输出就是下一层的输入，最后一层输出最终预测的结果。最后一层一般情况下是个softmax函数，实现把倒数第二层的输入映射成为一个[0,1]区间的集合，其中最大值就是预测的最终结果。所以数据通过模型在倒数第二层的输出就是数据通过模型提取得到的特征。如何才能得到倒数第二层的输出，我们只需要让最后一层把输入原封不动的输出就可以。具体实现就是删除掉模型原来的最后一层的函数:del model.fc ,然后让model.fc= lambda x:x。上面介绍的RestNet18模型最后一层输出的是一个长度为512的向量。

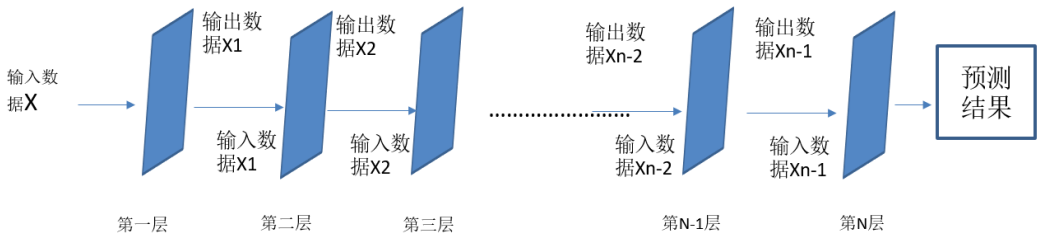


图 2.11 神经网络分层模型

2.7 GAN 网络

生成对抗网络[50]就是一个正方和敌方相互博弈的过程，相互促进，共同成长最后达到一个平衡。具体就是有一个生成器G和一个判别器D，G负责生成更加逼真的假数据来欺骗D，D的职责是通过不断训练提高判断假数据的能力从而可以非常准确的区分假数据和真数据，最终产生的结果是G生成的假数据，D无法判断真假，只能猜测，真假概率分别为50%。在机器学习中，我们可以利用GAN网络来生成一些不存在的数据，然后可以让模型做为真实数据使用从而达到数据投毒的目的。

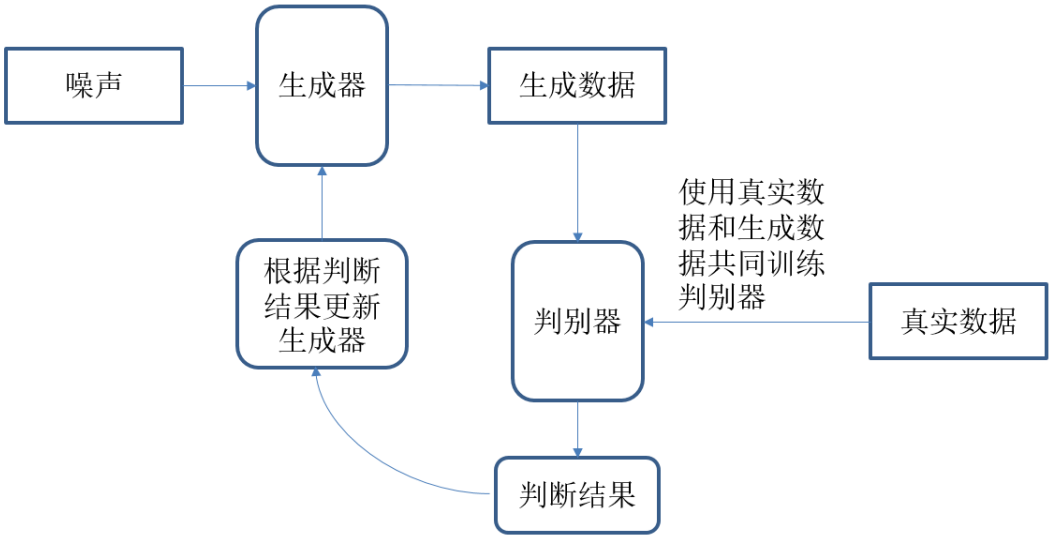


图 2.12 GAN 网络示意图

以生成图片为例，在训练过程中，生成器 G 的目标就是尽量生成真实的图片去欺骗判别器 D。而判别器 D 的目标就是尽量把生成器 G 生成的图片和真实的图片分别开来。这样，G 和 D 构成了一个动态的对抗过程。

GAN 模型的目标函数如下

$$\min_{G}\max_{D}V(D,G)=E_{x\sim pdata(x)}[\log D(x)]+E_{z\sim pdata(z)}[\log(1-D(G(Z)))] \quad (2.5)$$

公式的含义是，判别器 D 最大化自己正确分类真样本 x 和假样本 z 的概率，就是最大化 log D(x)和 log(1-D(G(Z)))，生成器 G 的目的就是要让判别器 D 识别生成数据为假数据的概率最小化，就是让 log(1－D(G(z)))的值最小化，同时也就是让 D 的损失最大化，从中可以清晰的感受到两个网络的对抗和博弈。在训练过程中 D 和 G 不是同时训练更新的，而是固定一个，更新另外一个网络的参数，交替迭代，提升自己的本领使得对方的犯错的概率不断增大，自己的正确概率不断增大，就是要削弱对手，增强自己，最终，生成器 G 在不断的训练下能估测出样本数据的分布，也就是生成的样本更加趋向于真实数据。

生成对抗网络在训练过程中，效果可能有如下几个过程，论文[50]画出的图如下：

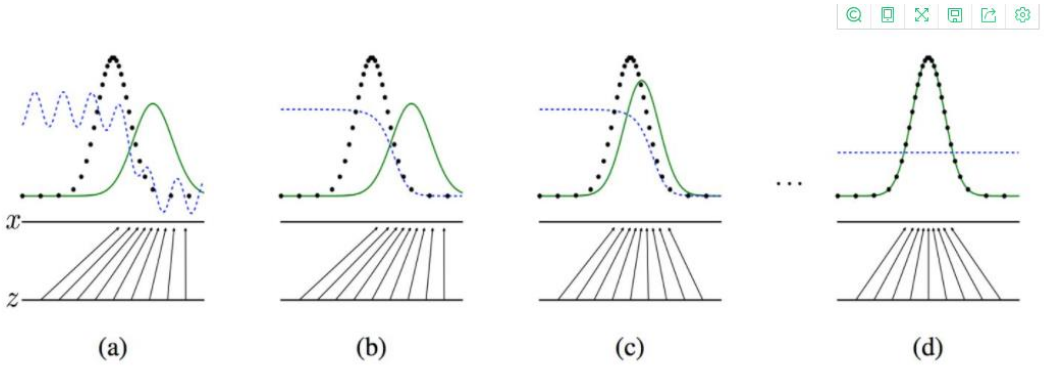


图 2.13 GAN 网络不同阶段效果图[50]

真实数据 x 的分布如黑色线所示，生成对抗网络生成的数据分布如绿色线所示，生成对抗网络生成的数据在判别器中的分布如蓝色线所示。

图a代表生成对抗网络刚开始训练可以基本的区分真实数据和生成数据但是有较大的波动，不够稳定。在图b中识别器D得到了训练，可以很好的区分出生成数据和真实数据。在图c中生成器G得到训练，生成的假数据的分布趋向于真实数据，我们发现绿色线越来越靠近黑色线。随着训练的持续，训练到最优的结果就是$P_{G(x)}=P_{data}$，就是$D(G(x))$的概率会趋近于0.5，这时候意味着识别器D无法正确区分真实数据和生成数据，只能随机判断，结果就是真假概率各为50%。图d就是最后理想的结果识别器和生成器都无法进一步改定，达到了一种平衡。

2.8 Pytorch 中梯度下降的实现

2.8.1 Pytorch

Pytorch的中文手册上指出Pytorch是torch的python版本，是由Facebook开源的神经网络框架，专门针对 GPU 加速的深度神经网络（DNN）编程。Torch 是一个经典的对多维矩阵数据进行操作的张量（tensor）库，在机器学习和其他数学密集型应用有广泛应用。与Tensorflow的静态计算图不同，pytorch的计算图是动态的，可以根据计算需要实时改变计算图。

2.8.2 梯度下降

在优化理论研究中，希望找到一个值使函数值达到全局最小。梯度下降法就是其中一种方法，这种方法当函数非常复杂，并且不能轻易使用数学代数求解函数时，这种方法依然可以很好的工作。当函数有很多参数时，一些其他方法不切实际，或者会得出错误答案，梯度下降法依然可以很好的工作。而且梯度下降法可以在数据不完善时，依然可以很好的工作。所以梯度下降法的提出促进了神经

网络的成熟和大发展。人工神经网络的训练主要采用梯度下降法，其计算过程中采用误差反向传播的方式计算误差函数对全部权值和偏置值的梯度。从而通过迭代更新的方式，不断更新模型参数，从而训练出一个满足要求的模型。

梯度下降方法的原理，大家都喜欢用下山的比喻来阐述。设想一下，在一个非常复杂、连绵不断的崇山峻岭，然后黑夜漫漫，伸手不见五指，你现在知道自己在一个山坡，想要到山底。你对周边一无所知，只能看到自己眼前 1 米左右的距离。这时候你能做的就是在这 1 米的范围内找到一个是下坡路的方向然后向下走，然后不断的走，不断地观察，在视野范围内寻找下坡路，直到最终走到坡底为止。梯度在这个例子中就是地面的坡度，你走的方向是最陡的坡度向下的方向。

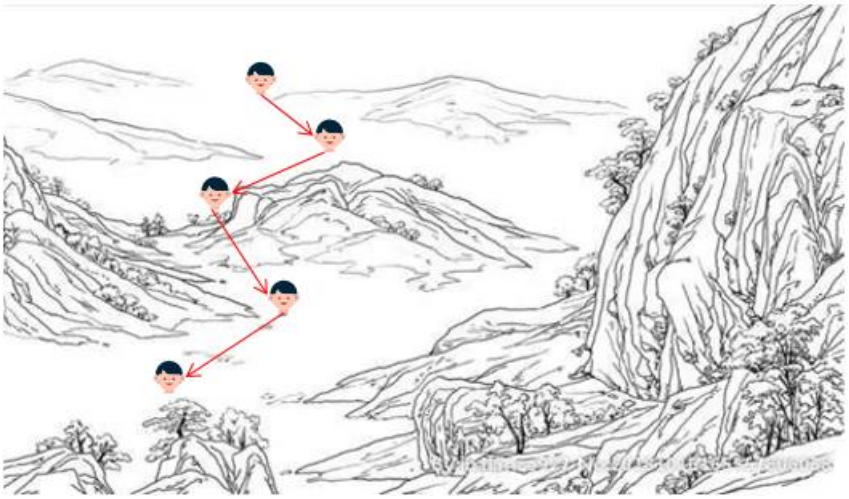


图 2.14 梯度下降示意图

现在让我们把这个复杂的地形看成是一个数学函数。梯度下降法使我们在不完全理解这个函数的情况下，从数学上对函数进行求解就可以找到最小值。梯度下降法让我们用步进的方式在局部寻找最优方式一小步一小步向前走，直到达成目标。

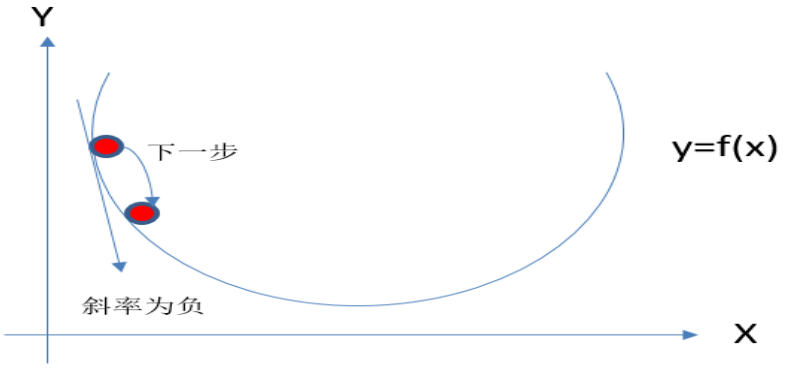


图 2.15 梯度下降斜率为负

假设有一个函数 $y=f(x)$,如图所示。我们希望找到 x,可以使 y 的值最小化。我们模拟梯度下降算法，先找一个起点，然后观察那个方向是向下的，也就是我们

标记的斜率为负的方向，因此我们沿着 X 轴向右走，我们稍微增加 x 的值，然后观察我们向实际最小值靠近了。

假设通过梯度下降，某一步走到了如下位置。此时，斜率为正，表明我们的通过步进方式越过了最小值，这时候需要向相反方向走，稍微减小 x 的值，更加靠近最小值。我们可以反复进行这样的操作，直到几乎不能改进为止，这样就得到了通过梯度下降法所能得到的最优解暨最小值。

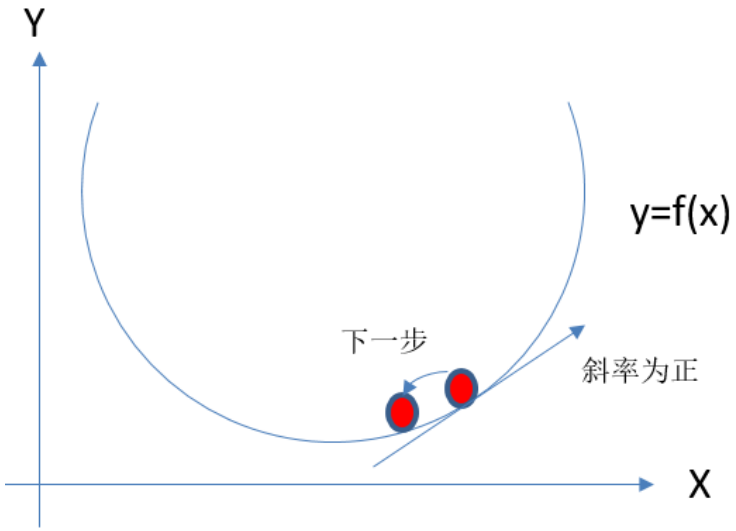


图 2.16 梯度下降斜率为正

我们要改变每一步的步幅，防止超调，这样可以避免在最小值附近来回震荡。假如我们距离最小值只有0.5米距离，但是我们采用2米的步长，那么由于向最小值的方向走的每一步都超过了最小值，我们就会错过最小值。这时我们就需要调节步长，与梯度的大小成比例，在接近最小值时，就需要使用小步长。通过图2.16所示，我们观察到需要往梯度的反方向增加x值，梯度为正的话，意味着函数值在增大，所以要减小x值，梯度为负的话，意味着函数值在减小，所以要继续增大x的值，使函数值减小。

2.8.3 Pytorch 中求梯度的两种情景

（1）标量求梯度

在 Pytorch 中，标量可以直接求导，举例如下：

2.6 标量求导举例

```
import torch
p = torch.randn(2,3, requires_grad=True)#定义了一个 2 行 3 列的张量，并且需要保留梯度信息
q = p*2+1
z = torch.norm(q-p)#得到 z 是一个具体的值，即是一个标量
z.backward()#backward 方法会根据链式法则自动计算出叶子节点的梯度值
print(p.grad)
```

下面我们以本文中处理图片为例，介绍了具体的以两张图片特征的差值为损失函数求导的过程，有助于后续理解生成有毒数据的过程。

2.7 使用标量求梯度,以本文中处理图片为例

```
from torch.autograd import variable as V
x = V(data, requires_grad=True)#用图片数据 data 定义一个需要求梯度的变量 x
currentImageFeat=Get_feature(x)[0]#求图片数据 x 通过模型得到的特征
loss=torch.norm(currentImageFeat - tarFeat)#求当前图片 X 的特征和目标图片特征的差值的 2 范数作为损失值。
loss.backward()#由于 loss 是个标量，可以直接求导
NewImage = data - learning_rate *x.grad  #通过得到的梯度 x.grad 来更新图片数据 data,从而得到新的图片 NewImage
```

（2）向量求梯度

向量不能直接对向量求导，要想实现向量对向量求导就要先设置一个权重系数 v，然后执行如下操作。这样对可以对向量 y 中的元素逐一求导，求导后分别乘以系数就是最终的值了。

2.8 向量求导举例

```
import torch
x = torch.tensor([1.0, 2.0, 3.0], requires_grad=True)
y = x*2
v = torch.tensor([0.1, 1, 0.01], dtype=torch.float)
y.backward(v)
print(x.grad)
# y=2x，y 对 x 求导为 2，因为 x 是向量，故为[2, 2, 2]，乘上系数为[0.2000, 2.0000, 0.0200]。
```

下面以本文中最后使用的图片处理为例介绍使用向量求导。

2.9 使用向量求梯度,以本文中处理图片为例

```
from torch.autograd import  variable as V
x = V(data, requires_grad=True)#用图片数据 data 定义一个需要求梯度的变量 x
currentImageFeat=Get_feature(x)[0]#求图片数据 x 通过模型得到的特征
loss= currentImageFeat – tarFeat #求当前图片 X 的特征和目标图片特征的差值，对于 ResNet18 来说特征就是一个长度为 512 的向量。
loss.backward(loss.clone())#由于 loss 是个长度为 512 的向量，backward()函数权重参数可以使用 loss 向量
NewImage = data - learning_rate *x.grad  #通过得到的梯度 x.grad 来更新图片数据 data,从而得到新的图片 NewImage
```

2.9 本章小结

本章主要介绍了本研究中需要了解的一些基础知识和实验中一些关键的实现手段。首先介绍了联邦学习的整体概念和分类，以及每种分类的一些关键点的实现方法，有利于读者阅读后续内容。后面依次介绍了后门攻击在联邦学习中的实现方式和手段，差分隐私在联邦学习中的应用，同态加密在联邦学习中的应用，ResNet18 模型的介绍，模型提取数据特征的原理和方法，GAN 网络的原理和实现过程，Pytorch 中梯度下降的实现。

第3章 基于特征的联邦学习后门攻击

联邦学习中各参与者不用共享自己的数据就可以联合其他参与者一起训练模型，从而可以得到更好的模型同时又保护了数据和隐私安全。在《数据安全法》等旨在保护数据和隐私安全的法律法规不断出台的大背景下，联邦学习日益受到重视和广泛应用。但是联邦学习在应用过程中也容易遭到攻击，其中一种就是后门攻击。研究攻击方式可以促进对于联邦学习自身安全的重视，促进联邦学习更加安全的发展。现有的后门攻击主要有两种一种是植入触发器，通过触发器激发后门攻击模型。一种是语义后门，通过改变具有特定语义的数据的标签，植入后门，让所有具有这种语义的数据都被识别为指定的标签。本章提出一种基于特征的后门攻击方法，根据两种不同场景提出了两种对应的后门攻击过程。主要思想就是选取具有指定标签的数据为目标数据，以目标数据的特征为目标处理原始数据，让生成的数据在特征上趋向于目标数据，这样就可以被模型识别为目标数据的标签类别，让数据在外观上保持不变，这样就可以防止人工检测等方式发现异常，从而增加后门攻击的隐蔽性。

3.1 攻击数据生成算法

基于特征的联邦学习后门攻击就是利用模型得到目标数据的特征记为 F_{target}，然后处理待分类数据使其在特征上趋向于 F_{target},在外观上基本保持不变，从而生成攻击数据，攻击数据在模型分类时被识别为目标数据的类别。联邦学习中常用的模型就是神经网络，神经网络模型在倒数第二层输出的结果是数据经过模型之后的特征 *f(x)*, *x* 为输入数据。基于特征的联邦学习后门攻击方法生成攻击数据的过程就是让公式(3.1)最小化的过程。

$$D = \arg\min_{X}(Dis(f(x), F_{target}) + Dis(x, D_{base})) \tag{3.1}$$

*Dis(x,y)*代表变量x和y之间的距离，D_{base}代表原始数据。*Dis(x,*D_{base}*)*使得攻击数据在外观上趋向于原始数据，*Dis(f(x),* F_{target}*)*使得攻击数据在特征上趋向于目标数据。通过公式（3.1）使得攻击数据D在特征空间上趋向于目标数据，在外观空间上趋向于原始数据。求解D的过程使用算法3.1[35]或算法3.2。两种算法都基于“forward-backward-splitting iterative procedure”算法[51]。

算法 3.1 文献[35]基于误差范数的攻击数据生成算法
输入：目标数据 t,原始数据 b，学习率 λ 初始 X：$X_0 \leftarrow b$ 定义 $L(x)=\| f(x)-f(t) \|_2$ for i=1 to Maxiter:# Maxiter 代表最大迭代次数 Forward step： $X_i^{'} = X_{i-1} - \lambda \nabla L(X_{i-1})$ Backward step: $X_i = (\lambda\beta b + X_i^{'})/(1+\lambda\beta)$ End for

算法 3.2 本文基于误差向量的攻击数据生成算法
输入：目标数据 t,原始数据 b，学习率 λ，分段步长 S 初始 X：$X_0 \leftarrow b$ 定义 L(x)=f(x)-f(t) for i=1 to Maxiter: # Maxiter 代表最大迭代次数 对于 $L(X_{i-1})$按照分段步长 S 分段，对于每一段向量求 2 范数，最后生成一个新的向量 $L^{'}(X_{i-1})$ Forward step： $X_i^{'} = X_{i-1} - \lambda \nabla L^{'}(X_{i-1})$ Backward step: $X_i = (\lambda\beta b + X_i^{'})/(1+\lambda\beta)$ End for

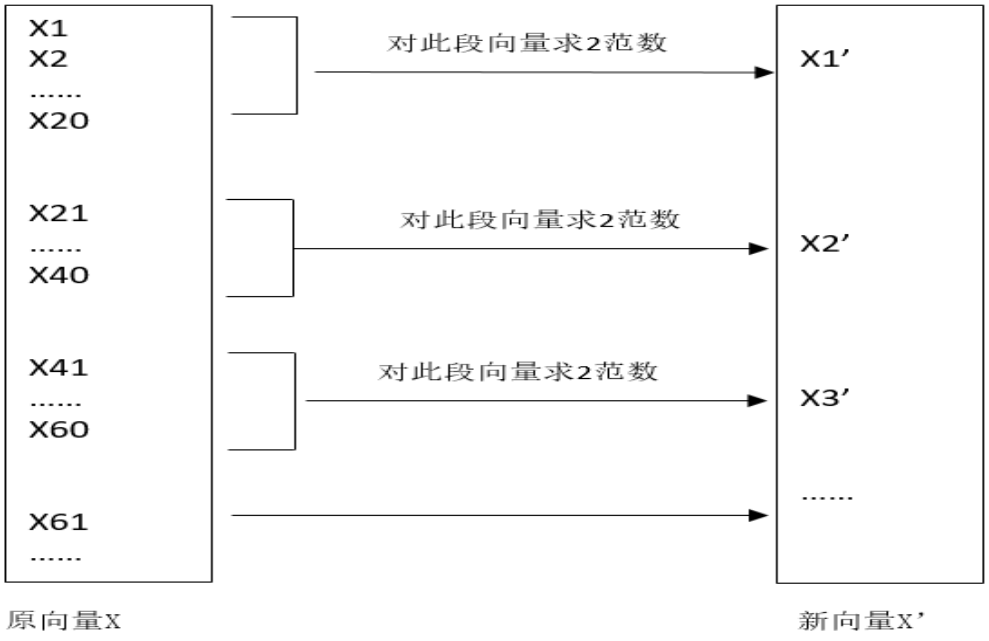


图 3.1 向量分段示意图

算法 3.1 基于误差范数的攻击数据生成算法定义的 L(x)是生成数据 x 和目标数据 t 两个特征向量差的 2 范数，用来度量 x 和 t 的距离。算法 3.2 基于误差向量的攻击数据生成算法定义的 L(x)是数据 x 和目标数据 t 两个特征向量的差,是一个

向量，然后把这个向量按照分段步长 S 分为几个子向量，对每个子向量求 2 范数，然后生成一个新的向量，作为梯度下降的损失值。图 3.1 以分段步长 20 为例演示了分段的过程。通过梯度下降使得 L(x)变小，从而使得生成数据 x 在特征上趋向于目标数据 t。在梯度下降的过程中算法 3.2 基于误差向量的攻击数据生成算法在相同训练条件下可以达到更优的效果。

3.2 基于特征的后门攻击方案

在联邦学习中的后门攻击可以根据恶意参与者所处的不同场景来来设置不同的攻击方案。

3.2.1 以恶意参与者参与正常训练的数据作为后门

在这种场景下，恶意参与者拥有需要攻击的目标类型的数据。比如恶意参与者拥有标签为 2 的鸟类数据，然后希望任意的图片被分类为标签为 2 的鸟类数据，这时恶意参与者在训练过程中只需要以正常的方式参与训练，不需要投入任何后门数据，不需要做任何恶意行为，这样恶意参与者就相当于一个正常的参与者，参与训练的概率极大提高，也更加隐蔽。训练完成后得到全局训练模型，恶意参与者以自己拥有的标签为 2 的鸟类数据作为后门，然后以这些图片为目标图片用算法 3.1 或算法 3.2 来处理原始图片，最终生成攻击图片。攻击方案如图 3.2 所示。

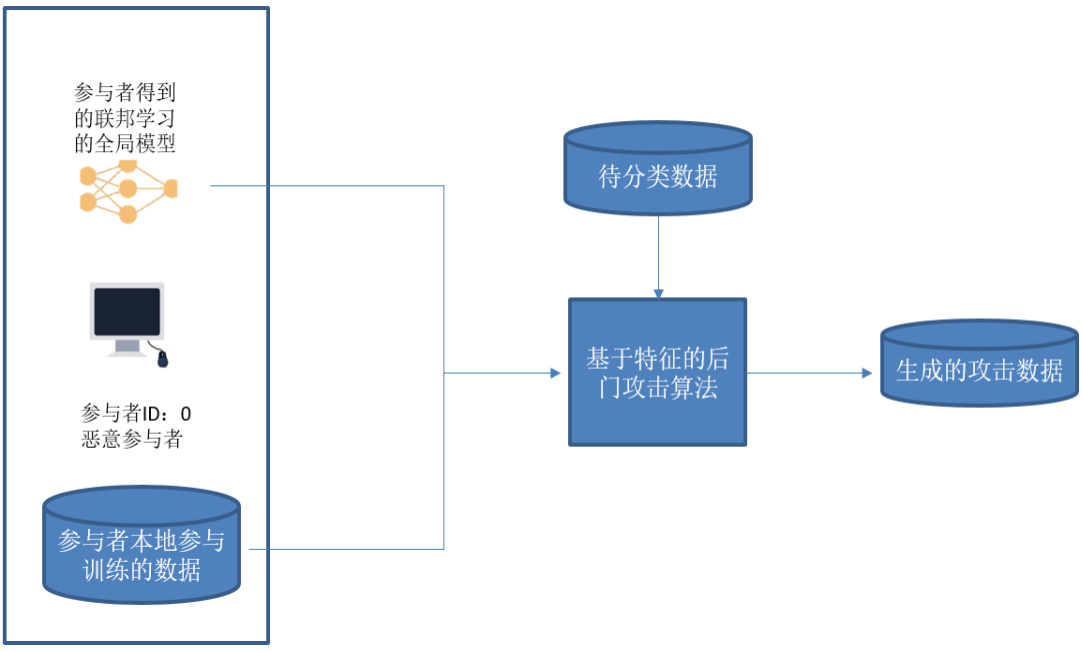


图 3.2 以恶意参与者参与正常训练的数据为后门的攻击方案

3.2.2 恶意参与者植入后门数据

在这种场景下，恶意参与者不拥有需要攻击的目标类型的数据。比如恶意参

与者不拥有标签为2的鸟类数据，同时又希望任意的图片被分类为标签为2的鸟类数据，这时需要在训练过程中选择一些其他的图片作为后门图片参与训练，标签设置为鸟类数据的标签2。在训练完成后，以投入的后门图片为目标图片用算法3.1或算法3.2处理原始图片，最终生成攻击图片。攻击方案如图3.3所示。

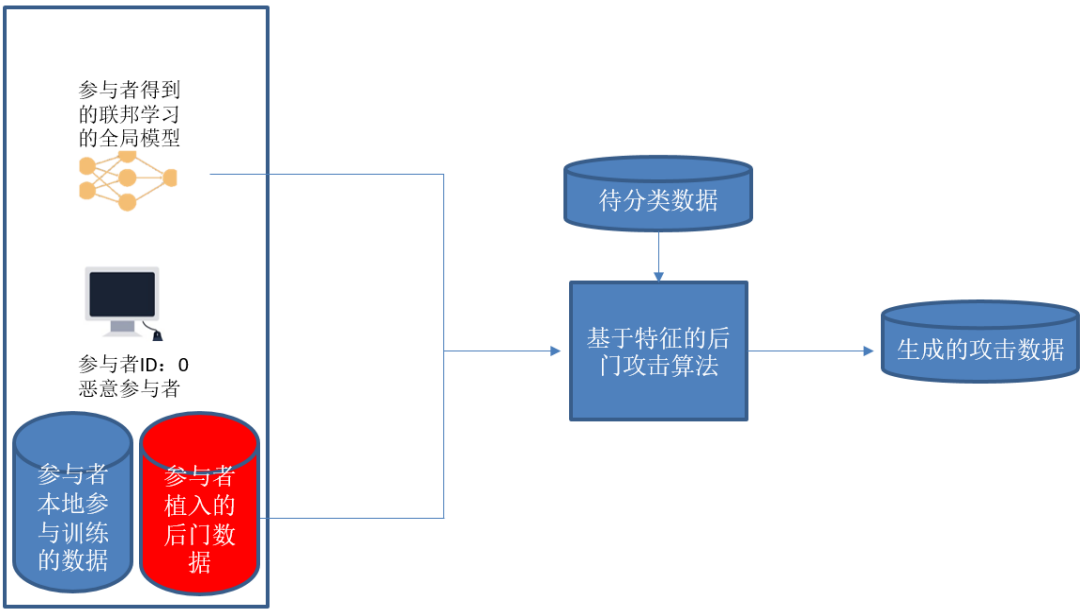


图 3.3 以恶意参与者植入的后门数据为目标的攻击方案

3.3 使用目标数据的方法

有了目标数据，如何使用这些目标数据，可以有以下两种方式：

第一种是取所有目标数据的特征的平均值作为目标特征，称为目标特征值平均法。依次提取所有目标数据的特征值，特征值就是一个向量，然后对所有数据的特征向量求和取平均值得到一个最终使用的目标特征向量。

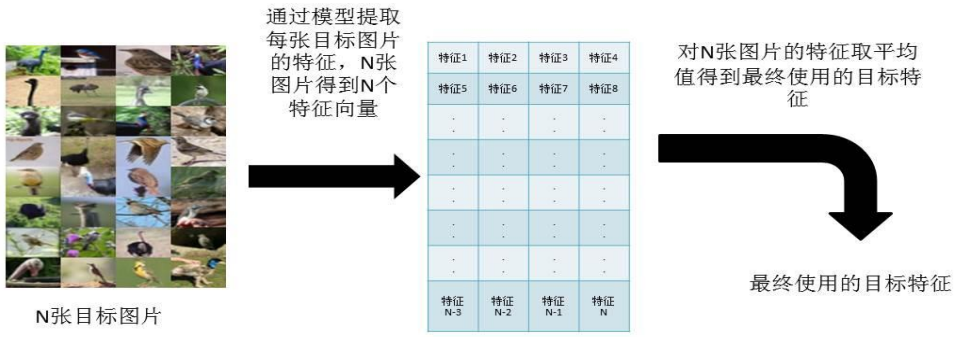


图 3.4 通过平均值法求得最终使用的目标特征

第二种是在所有目标数据中找和原始数据最相近的数据作为目标数据，称为目标数据相近法。选择最相近的目标数据有三种方法，分别是数据的特征差值的 2 范数最小，数据的原始值的差值的 2 范数最小，数据的特征差值的 2 范数和数

据的原始值的差值的 2 范数之和最小。

3.4 实验

实验环境：阿里云桌面、Inter(R) Xeon(R)KVM CPU @2.50GHz 2.50GHz 处理器，8GB 内存，64 位 Windows 操作系统，Python 程序语言，开发框架 PyTorch，数据集 CIFAR-10,模型 ResNet18。联邦学习模型使用典型的服务器-客户端模式，10 个参与者其中 1 个为恶意参与者，1 个模型参数服务器。

3.4.1 以恶意参与者参与正常训练的数据作为后门

以恶意参与者参与正常训练的数据作为后门场景下，恶意参与者拥有待攻击目标类型的数据，恶意参与者只需要正常参与联邦学习训练，不需要投入有毒数据。实验中把 CIFAR-10 数据集中的训练数据等分为 10 份，设置 10 个参与者，参与者的 id 依次为 0 到 9，其中设置 id 为 0 的参与者是恶意参与者，每一个参与者拥有一份数据。使用模型 ResNet18 进行 30 轮训练之后得到一个识别率为 87.94%的模型记为 Model1，其中训练时图片尺寸为 64*64。假设 id 为 0 的恶意参与者需要攻击标签为 2 的鸟类数据，就是让任意图片被识别为标签为 2 的鸟类数据，模型如图 3.5 所示。

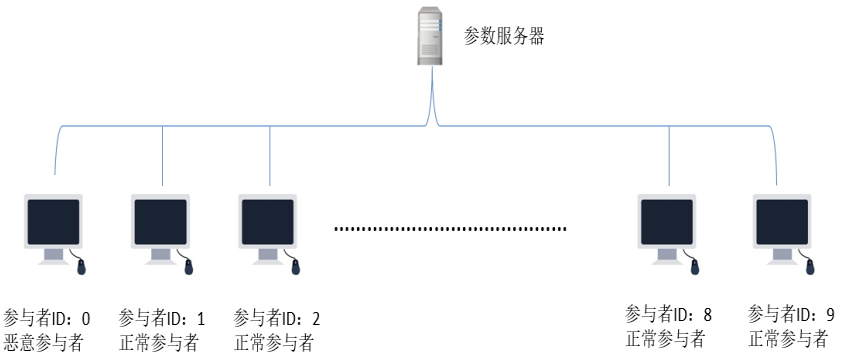


图 3.5 训练模型

（1）目标特征值平均法

从恶意参与者拥有的图片中选择部分标签为 2 的鸟类图片，实验中我们选择了 92 张鸟类图片，这 92 张鸟类图片可以被模型 Model1 百分之百的正确识别。通过 Model1 取得每一张图片的特征向量，然后取 92 张图片特征向量的平均值得到目标特征向量 F_{target}。利用算法 3.1，算法 3.2 分别使用 50 张其他类别的测试图片在 500 轮训练条件下生成 50 张攻击图片，攻击图片生成流程如图 3.6 所示。CIFAR-10 图片集共有十个类别，分别是：label0:airplane, label1:automobile, label2:bird, label3:cat, label4:deer, label5:dog, label6:frog, label7:horse, label8:ship,

label9:truck。选择 label0，label1，label3，label4，label5,分别按照上述过程分别使用两种方法生成攻击图片，两种算法攻击成功率如表 3.1 所示，两种算法不同类别图片攻击成功率对比如图 3.7 所示。可以看出算法 3.2 在同等条件下后门攻击效果优于算法 3.1。生成的攻击图片如图 3.8 所示。十个类别按照算法 3.2 生成的攻击数据的攻击率如图 3.9 所示。从图中可以看出攻击率平均在 50%左右，对于模型的危害还是比较大。

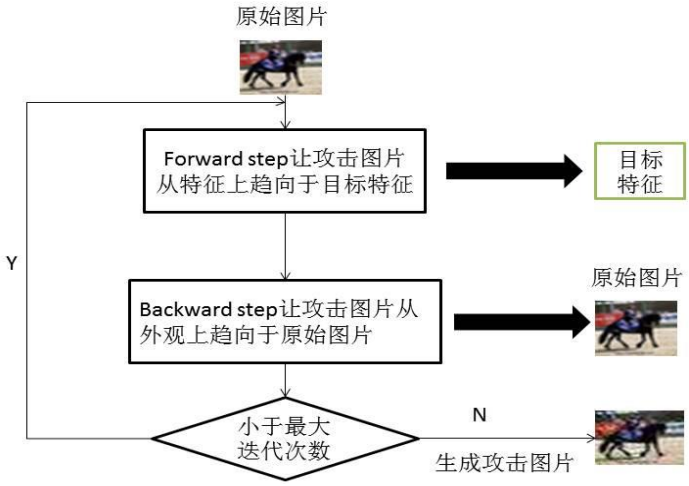


图 3.6 生成攻击图片流程

表 3.1 两种不同算法攻击成功率

目标图片类别	原始图片个数	算法	攻击成功率
label0：airplane	50 张	算法 3.2	47%
label1：automobile	50 张	算法 3.2	86%
label2：bird	50 张	算法 3.2	55%
label3：cat	50 张	算法 3.2	34%
label4：deer	50 张	算法 3.2	60%
label5：dog	50 张	算法 3.2	30%
label0：airplane	50 张	算法 3.1 文献[35]	22%
label1：automobile	50 张	算法 3.1 文献[35]	36%
label2：bird	50 张	算法 3.1 文献[35]	34%
label3：cat	50 张	算法 3.1 文献[35]	8%
label4：deer	50 张	算法 3.1 文献[35]	16%
label5：dog	50 张	算法 3.1 文献[35]	6%

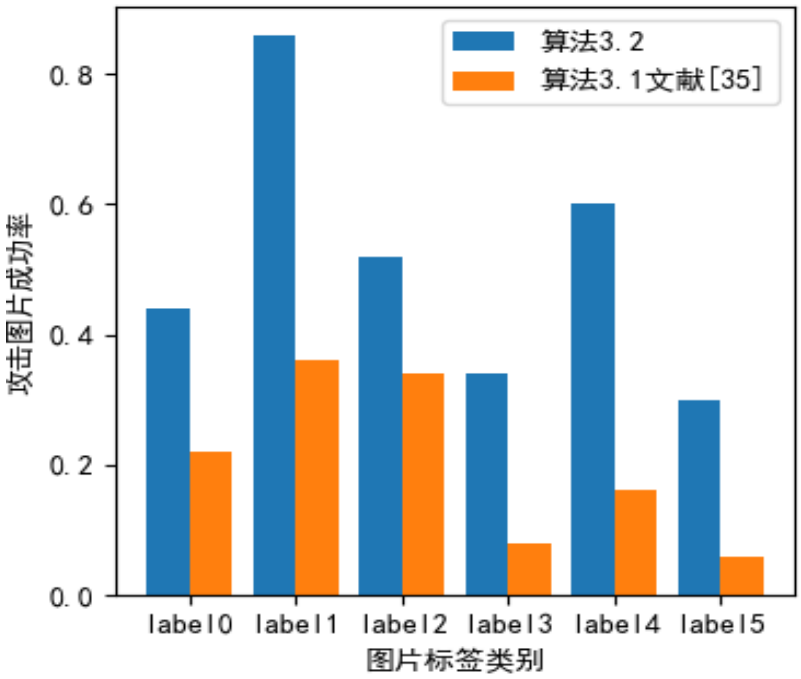
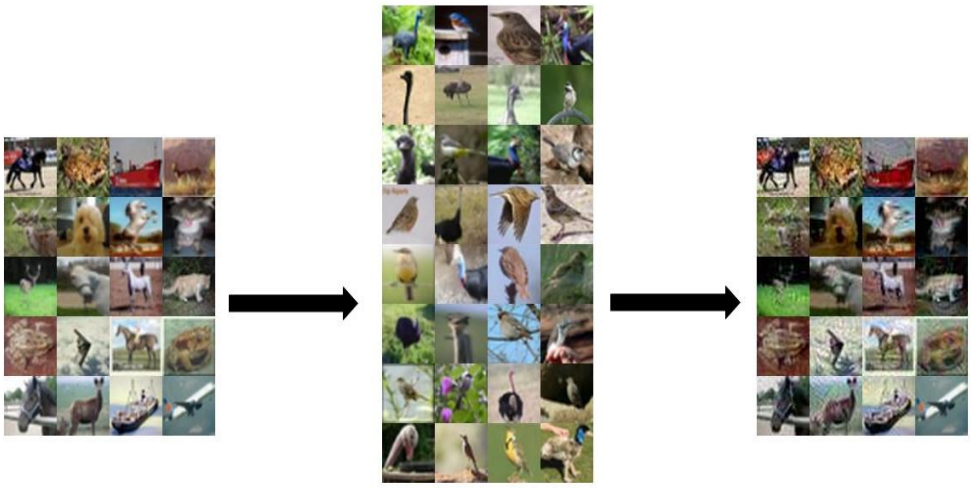


图 3.7 两种算法攻击成功率对比图



原始图片组　　目标图片组　　攻击图片组

原始图片通过处理程序以目标图片特征的平均值为目标进行处理,最终得到攻击图片,攻击图片在特征上趋向于目标图片,会被大概率的识别为目标图片的类别鸟类,在外观上趋向于原始图片。

图 3.8 攻击图片效果示意图

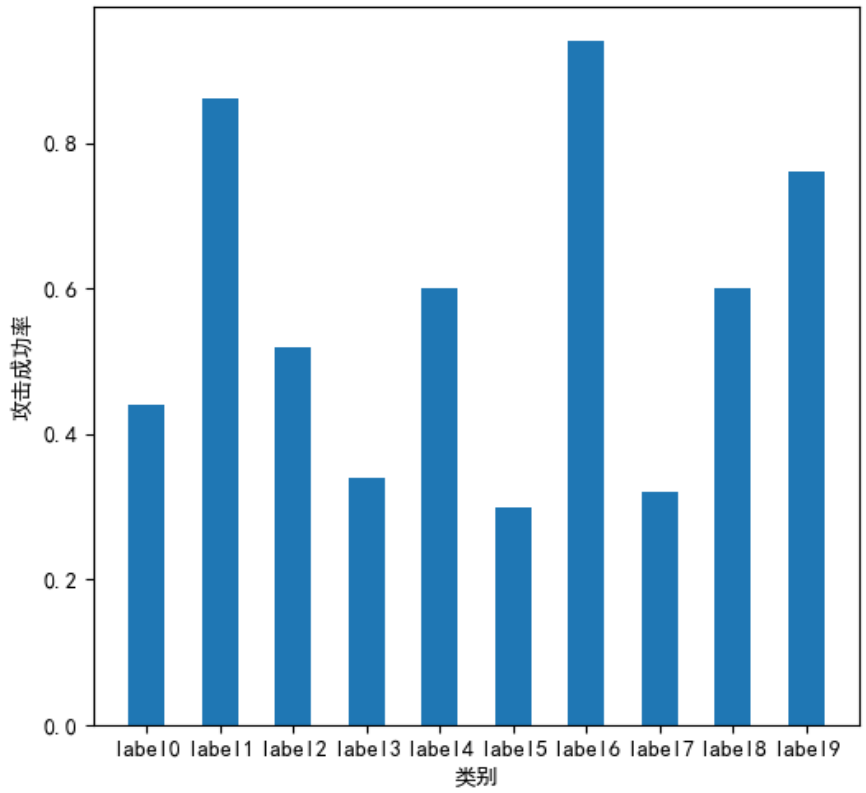


图 3.9 算法 3.2 生成的 10 种类别图片的攻击成功率

（2）目标数据相近法

恶意参与者从拥有的数据中选择部分标签为2的鸟类图片作为目标数据，和上面实验一致选择92张鸟类图片，这92张图片可以被模型Model1百分之百的正确识别。随机取50张非鸟类原始图片，对每一张原始图片在92张目标图片中选择一张最相近的作为目标图片进行处理，以算法3.2生成攻击图片。选择最相近的方式有三种：1.两张图片特征最相近2.两张图片原始值最相近3.两张图片的特征和原始值距离之和最相近。三种方式以2范数表示距离。对十种类别图片分别按照上述方法生成攻击图片。在采用算法3.2的情况下，三种相近法生成的攻击图片攻击成功率和目标特征值平均法生成的攻击图片的攻击率对比图如图3.10所示。通过图示，我们可以看出在以恶意参与者参与训练的正常数据作为后门时，目标特征值平均法成功率更高，目标数据相近法成功率整体偏低一些。在目标数据相近法中原始值相近的方法成功率更高。所以在此场景下优先使用目标特征值平均法，其次使用原始值相近的相近法。作为攻击者可以交替使用这些方法，寻找最优的攻击效果。

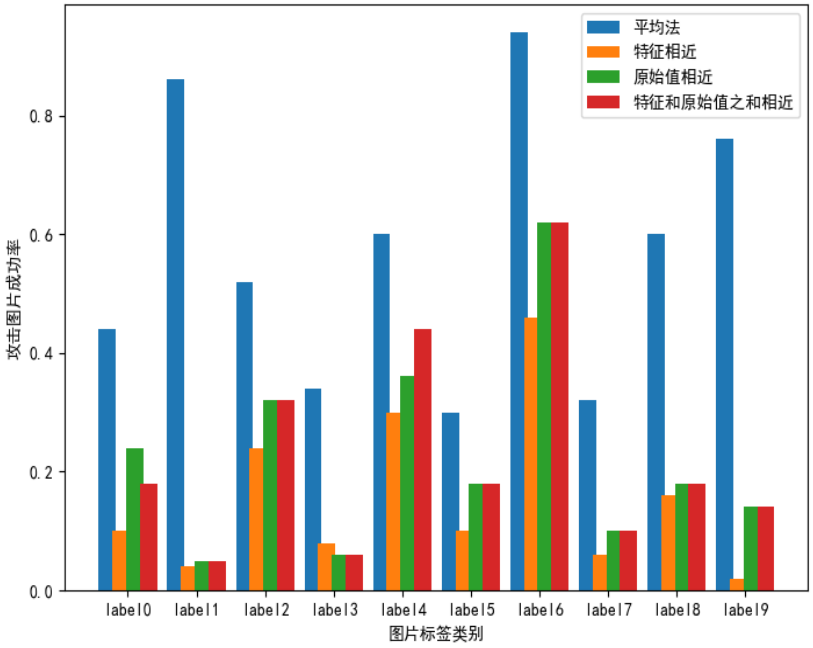


图 3.10 场景 1 下使用目标图片的四种方法的攻击成功率

3.4.2 恶意参与者植入后门

此种场景下，恶意参与者没有待攻击目标类型的数据，需要在联邦学习过程中植入后门数据。

（1）植入后门数据

共训练50轮，每轮10个参与者，参与者ID依次为0到9，每一个参与者去掉和自己id号相同的类别训练数据。ID为0的参与者作为恶意参与者去掉类别为0的训练数据，然后选取100张CIFAR-100数据集中类别为30的图片作为CIFAR-10中类别为0的后门数据参与训练，图片大小设置为64*64，得到一个模型记为Model2，识别率为87.48%，对100张植入后门图片识别率为95%。每一个参与者在本地训练完成后向参数服务器提交自己的模型参数更新差值 Δ 。

$$\Delta = T_i^{t+1} - G^t \tag{3.2}$$

T_i^{t+1}代表第i个参与者t+1轮训练后获得的本地模型参数，G^t代表第t轮时全局模型训练参数。50轮训练，10个参与者，取每轮每个参与者模型参数更新差值的2范数得到如图3.11的分布图，可以看出模型参数更新值正常参与者和恶意参与者非常接近。我们的实验中由于恶意参与者也是和其他参与者一样拥有相同数量的图片4500张，然后投入100张后门图片，后门图片占比很小，所以恶意参与者和正常的参与者的模型参数更新差值相差很小不易区分。但是如果恶意参与者的数据和正常参与者的数据相差较大时，恶意参与者和正常的参与者的模型参数更新差值相差就会比较大，就会出现离群点。但是联邦学习的特点又导致了很多时候各

参与者的数据是非独立同分布的，离群点检测并不一定适用。

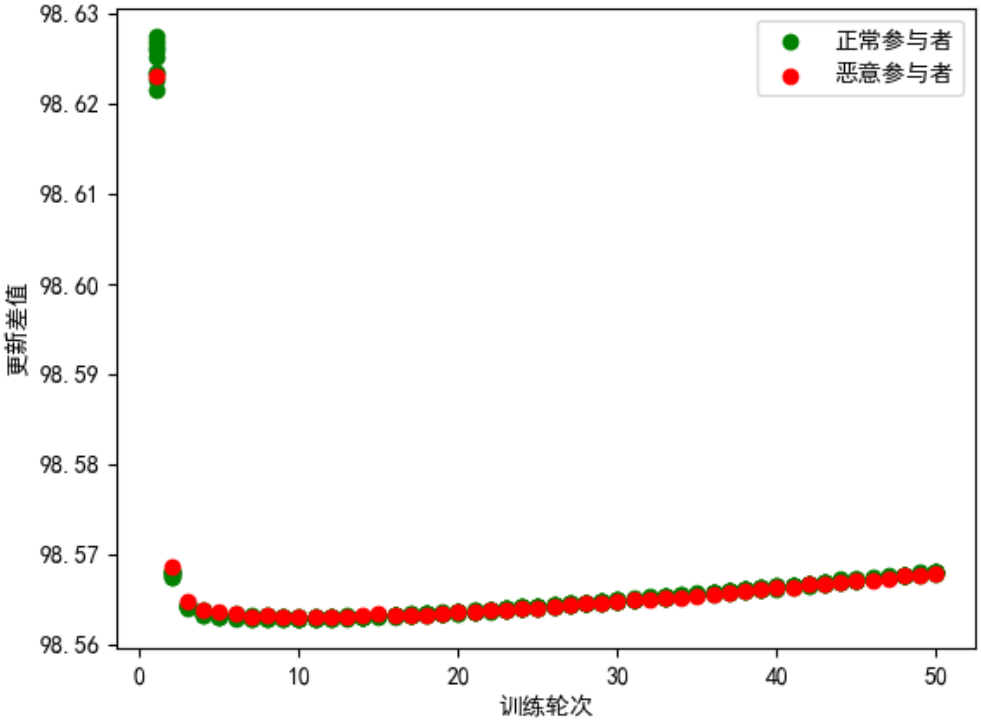


图 3.11 参与者模型参数更新差值分布图

（2）生成攻击数据

第一种从植入的 100 张后门图片中选择 95 张能够 100%被正确识别的图片为目标数据，采用目标特征值平均法生成攻击数据，50 张攻击数据成功率 32%。

第二种以选择的 95 张后门图片为目标数据，在目标图片中寻找与原始图片最相近的图片作为目标图片来生成攻击图片，使用三种选择相近目标数据的方法各生成 50 张攻击图片，成功率最高为 48%。从实验数据来看，植入后门时，选择目标数据相近法中的图片原始值最相近的方法成功率更高一些。

表 3.2 场景 2 下三种相近目标数据方法攻击成功率

目标图片类别	目标图片个数	原始图片个数	相近算法	攻击成功率
label0：airplane	95 张	50 张	两张图片特征最相近	30%
label0：airplane	95 张	50 张	两张图片原始值最相近	48%
label0：airplane	95 张	50 张	两张图片的特征距离和原始值距离之和最相近	48%

3.5 在不同尺寸图片集合下的实验

图片的尺寸大小会影响模型的识别率，上面的实验中我们使用的图片尺寸是64*64，为了研究不同尺寸大小图片对后门攻击率的影响，我们在实验中将使用32*32和128*128两种图片尺寸，分别进行模型训练和后门攻击。

模型的识别率指的是模型正确分类测试数据的准确率，攻击率指的是训练中投入的有毒图片被模型按照指定标签准确分类的准确率。从图3.12看出图片尺寸变大后模型的识别率得到很大提升，而且更加快速的收敛。可以看出模型在前10轮基本是垂直上升，识别率快速增加，尺寸为64和128的两个训练模型几乎在前二十轮时已经收敛了，后面的增长非常缓慢。图片尺寸增大之后，模型的攻击率也快速提升，也就说尺寸的增大既会提升模型识别率，同时也会让攻击成功率大大提升。

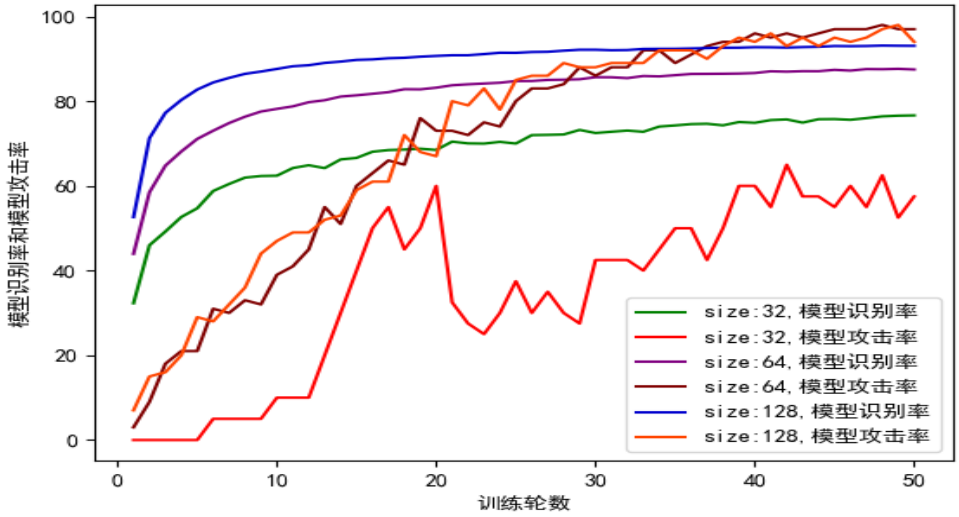


图 3.12 不同尺寸模型的识别率和攻击率

将尺寸为32的CIFAR-10图片数据集生成的模型记为model32，尺寸64的CIFAR-10图片数据集生成的模型记为model64,尺寸128的CIFAR-10图片数据集生成的模型记为model128。三种模型都训练50轮， ID为0的参与者为恶意参与者，训练中恶意参与者投入100张后门图片。训练完成后，使用训练得到的模型从100张后门图片中选取可以被正确识别的图片，model32选出了70张图片，modle64选出了95张图片，model128选出了94张图片。然后以选出的图片为目标图片在三种模型下分别使用目标特征值平均法，目标数据相近法来生成攻击图片，然后分别测试攻击图片的攻击成功率。从表3.3中可以看出，model64和model128模型下的攻击率更高，有相对明显提升，适当的增大模型图片尺寸可以提高训练的效率，让模型的识别率显著提升，同是也会提高本方案后门攻击的成功率。

表 3.3 不同尺寸模型场景 2 下的攻击成功率

目标图片类别	目标图片个数	原始图片个数	算法	攻击成功率	模型/识别率
Label 0：airplane	70 张	50 张	特征值平均法	3%	Model32 76.2%
Label 0：airplane	70 张	50 张	两张图片特征最相近	8%	Model32 76.2%
Label 0：airplane	70 张	50 张	两张图片原始值最相近	8%	Model32 76.2%
Label 0：airplane	70 张	50 张	两张图片的特征距离和原始值距离之和最相近	8%	Model32 76.2%
Label 0：airplane	95 张	50 张	特征值平均法	32%	Model64 87.48%
Label 0：airplane	95 张	50 张	两张图片特征最相近	30%	Model64 87.48%
Label 0：airplane	95 张	50 张	两张图片原始值最相近	48%	Model64 87.48%
Label 0：airplane	95 张	50 张	两张图片的特征距离和原始值距离之和最相近	48%	Model64 87.48%
Label 0：airplane	94 张	50 张	特征值平均法	44%	Model128 93.06%
Label 0：airplane	94 张	50 张	两张图片特征最相近	36%	Model128 93.06%
Label 0：airplane	94 张	50 张	两张图片原始值最相近	52%	Model128 93.06%
Label 0：airplane	94 张	50 张	两张图片的特征距离和原始值距离之和最相近	52%	Model128 93.06%

同时我们也选取了 label1,lable2,label3,label4 按照以恶意参与者正常训练数据作为后门的方式，每个类别选取 100 张对应的目标图片生成了三种模型下的攻击图片，对应的攻击成功率如图 3.13 所示。我们可以看出，model128 的攻击成功率比较高，而且也比较稳定，model64 和 model32 相对较低，可以看出训练图片尺寸较大的模型遭受攻击的概率也会增大。

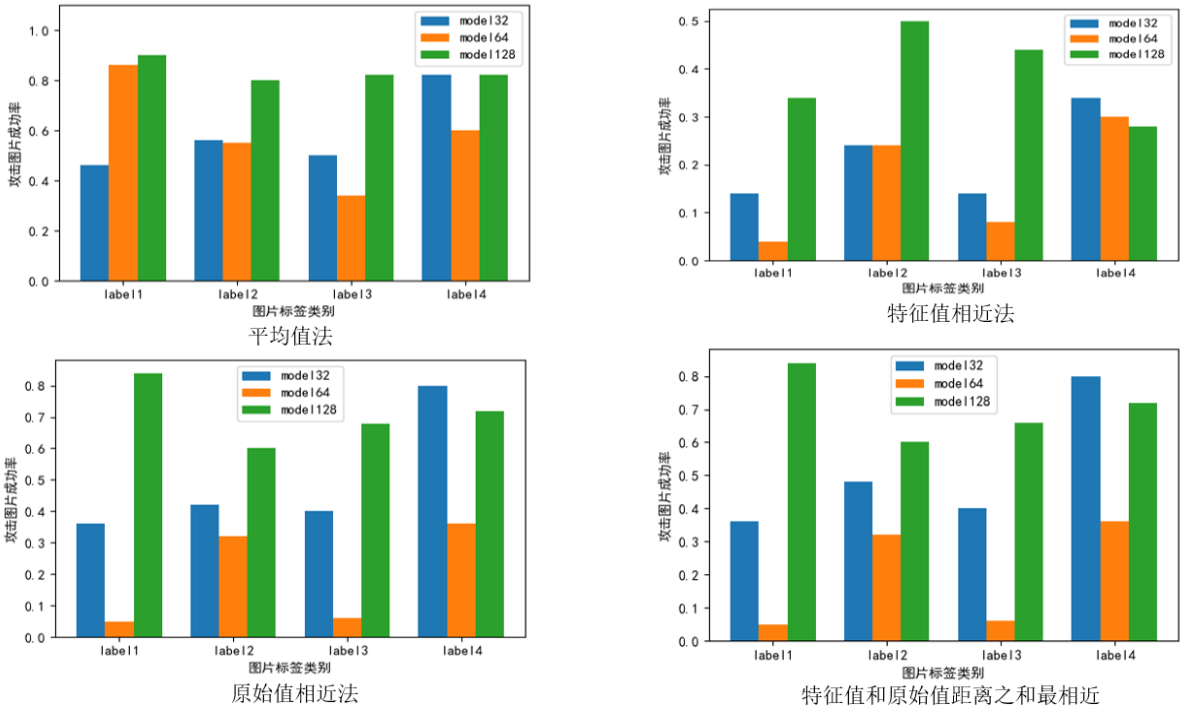


图 3.13 三种模型下攻击图片成功率对比图

3.6 本章小结

本章详细阐述了两种不同的前向后向分裂迭代算法，分析了二者的不同之处。然后介绍了在上述算法基础上的基于特征的后门攻击方法，首先介绍了两种不同场景下的对应的后门攻击方法，然后详细介绍了生成攻击数据的目标特征值平均法和目标数据相近法两种方法，相近法中又包括了三种：特征值最相近，原始图片值最相近，图片的特征距离和原始值距离之和最相近。在介绍完这些算法之后，展示了实验结果。实验结果表明以恶意参与者正常训练的数据作为后门是，攻击成功率在 50%左右。在恶意参与者植入后门图片场景下，以植入的后门图片为目标来处理原始图片从而生成攻击图片，这种方式下攻击成功率相对不高，后续需要研究加入注意力机制之后能否提高攻击成功率。最后讨论了通过不同尺寸图片训练得到的模型在模型识别率，生成攻击图片的攻击率方面的不同，通过实验展示了具体的结果。在实验结果中生成的攻击图片在 model64 和 model128 中的攻击率相对高一些，在 modle32 相对低一些，适当增大图片尺寸可以有效的提高模型的识别率，但是也会提高攻击的成功率。

第4章 基于生成对抗网络和特征的联邦学习后门攻击方法

生成对抗网络 GAN 就是一个正方和敌方相互博弈的过程，相互促进，共同成长最后达到一个平衡。具体就是有一个生成器 G 和一个判别器 D，G 负责生成更加逼真的假数据来欺骗 D，D 的职责是通过不断训练提高识别假数据的能力从而可以非常准确的区分假数据和真数据，最终产生的结果是 G 生成的假数据，D 无法判断真假，只能猜测，真假概率分别为 50%。在机器学习中，我们可以利用 GAN 网络来生成一些不存在的数据，然后可以让模型作为真实数据使用从而达到数据投毒的目的。

使用生成对抗网络的前提条件是判别器 D 必须要能够得到真实数据从而来训练判别器 D，提升判别器 D 识别真假数据的能力。但是联邦学习中，各参与者不用共享各自的本地数据，也就对应着各参与者无法获取其他参与者的数据，没有真实数据，就无法训练判别器 D。但是联邦学习的特点使得恶意参与者在训练过程中可以得到全局模型。全局模型是由所有参与者的真实数据训练之后经全局聚合之后得到的。以全局模型作为生成对抗网络中的判别器 D 就相当于在所有参与者的真实数据上训练判别器。有了判别器 D，恶意参与者就可以训练生成器 G，生成更加真实的假数据，从而用假数据攻击模型。

4.1 整体方案

联邦学习中，恶意参与者在整体训练结束后得到全局模型。全局模型是由所有参与者的真实数据训练之后经全局聚合之后得到的。以全局模型作为生成对抗网络中的识别器D就相当于在所有参与者的真实数据上训练出判别器。我们以全局模型作为判别器D来训练生成器G，让G生成指定标签的生成数据。在生成数据中选取可以被全局模型正确识别为指定类别的数据，然后以这些数据作为目标数据，然后以基于特征的后门攻击方法生成攻击数据，从而达到模型攻击的目的。已有的基于生成对抗网络GAN的后门攻击都是在生成数据后把这些数据作为触发器植入训练模型，经过训练之后得到植入了后门的全局模型。本方案在通过生成对抗网络GAN得到生成数据后，通过基于特征的后门攻击方法生成攻击数据，攻击数据从特征上趋向于生成数据，在外观上趋向于原始数据，从而以更加简单的方式攻击模型。攻击方案如图4.1所示。

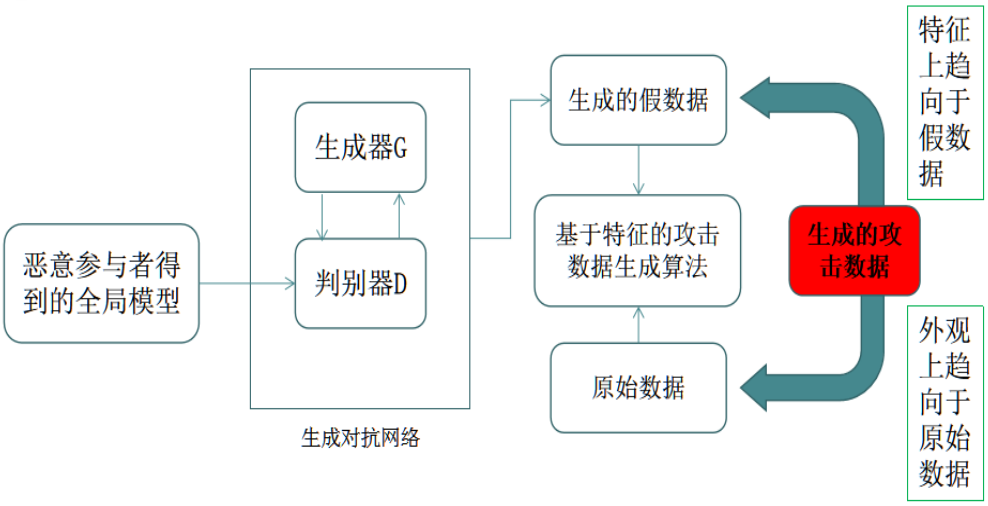


图 4.1 基于生成对抗网络和特征的后门攻击方案

4.2 使用生成对抗网络生成数据

生成对抗网络由判别器D和生成器G组合而成，通过判别器D和生成器G之间的博弈，从而达到让生成器G产生让识别器D无法判断是真还是假的生成数据。联邦学习中各参与者不共享数据，只用本地数据训练本地模型，然后把梯度信息或者模型参数提交给协调服务器，由服务器对所有参与者提交的参数信息进行安全聚合，然后把聚合后的信息下发给所有参与者，以这种方式达到联合训练模型的目的。GAN网络中要训练识别器D需要两部分数据一部分是真实数据，让D学习真实数据的分布，另一部分是生成器G产生的假数据，让D学习假数据的分布，通过不断的交叉训练最终产生的假数据分布趋向于真实数据的分布从而使得判别器无法判断。但是在联邦学习中每个参与者看不到别的参与者的数据，更不用说全部的数据，没有真实数据，如何来训练判别器D。联邦学习的特点使得每个参与者可以得到全局模型，而全局模型是在所有参与者真实数据上训练之后聚合而来，就是一个天然的判别器。有了判别器D我们就可以正常的开始训练GAN网络从而来产生假数据。在生成假数据时，固定判别器D，指定要生成的数据类别。首先产生噪声，通过生成器G生成假数据，使用判别器评估准确率是否达标，如果不达标，计算损失值，然后通过梯度下降更新生成器G的参数，反复迭代直到生成器G能够生成准确率达标的生成数据，输出生成数据。这些生成数据能够被模型准确识别为对应类别，以这些生成数据做为目标数据来处理原始数据从而生成攻击数据。

下面我们详细介绍一下生成网络 G 的结构。

4.1 生成器 G

```
class netG(nn.Module):
def __init__(self):
        super().__init__()
        self.linear = nn.Linear(nz, 64*2*2)
        self.layer1 = nn.Sequential(
            BasicBlock(64),
            nn.UpsamplingNearest2d(scale_factor=2),
            BasicBlock(64),
            nn.UpsamplingNearest2d(scale_factor=2),
        )
        self.layer2 = nn.Sequential(
            BasicBlock(64),
            nn.UpsamplingNearest2d(scale_factor=2),
        )
        self.layer3 = nn.Sequential(
            BasicBlock(64),
            nn.UpsamplingNearest2d(scale_factor=2)
        )
        self.layer4 = nn.Sequential(
            BasicBlock(64),
            nn.UpsamplingNearest2d(scale_factor=2)
        )
        self.Conv = nn.Sequential(
            BasicBlock(64),
            nn.BatchNorm2d(64),
            nn.ReLU(True),
            nn.Conv2d(64, 3, kernel_size=3, padding=1, stride=1),
            nn.Tanh()
        )
    def forward(self, z):
        x = self.linear(z)
        x = x.view(batch_size,64,2,2)
```

```
        x = self.layer1(x)
        x = self.layer2(x)
        x = self.layer3(x)
        x = self.layer4(x)
        x = self.Conv(x)
        return x
```

下面我们介绍一下判别器D，D使用的是Pytorch中实现的ResNet18模型，所以直接调用Pytorch中的模型，此处不再赘述。

利用判别器 D 和生成器 G 生成假数据的训练过程中先初始化生成器 G，然后用联邦学习的全局模型初始化判别器 D，然后开始多轮训练。每轮训练中先生成随机噪声，噪声通过生成器 G 生成假数据，假数据通过判别器 D 得到类别结果，然后和目标类别结果计算误差得到损失，然后执行损失的反向传播，更新模型，通过反复训练得到可以被识别为指定标签的数据。

4.2 GAN 网络训练过程

```
    netG = netG().to(device)#定义生成网络实例
    netG.apply(weights_init)#初始化参数
    netD = torch.load('testmodle/1.pt')#1.pt 是联邦学习的全局模型，通过 load 加载把全局模型赋给了判别器 D。
    fake_label=0#设置假数据指定的标签
    label = torch.full((opt.batchSize,), fake_label,
                dtype=torch.long, device=device)#设置长度为 opt.batchSize 的标签数组
    for epoch in range(opt.epoch):#循环 opt.epoch 轮次
        for i in range(20):#每个轮次循环 20 次
            # 固定鉴别器 D，训练生成器 G
            print("start {}轮{}次".format(epoch,i))
            optimizerG.zero_grad()
            noise = torch.randn(opt.batchSize, opt.nz)#生成随机噪声
            noise = noise.to(device)
            fake = netG(noise)  # 生成假数据
            output = netD(fake)#通过判别器判别真假
            pred=output.data.max(1)[1]
            correct=0
            correct += pred.eq(fake_label).cpu().sum().item()
            lossG = torch.nn.functional.cross_entropy(output,label)#计算损失值
            lossG.backward()#反向梯度传递
```

```
            print("loss is {}".format(lossG))
            optimizerG.step()#修改模型参数
```

4.3 以生成的假数据为目标数据生成攻击数据

通过GAN网络生成可以被联邦学习全局模型正确识别的假数据，以生成的假数据为目标生成攻击数据。以生成的假数据为目标数据生成攻击数据，有两种使用目标数据的方法。

一种是特征值平均法。通过GAN网络生成指定标签的假数据，提取每一张假数据通过联邦学习全局模型后的特征，然后取这些特征的平均值，以特征的平均值为目标特征利用基于特征的后门攻击方法生成特征上趋向于假数据，外观上基本保持不变的攻击数据，达到攻击模型的目的。

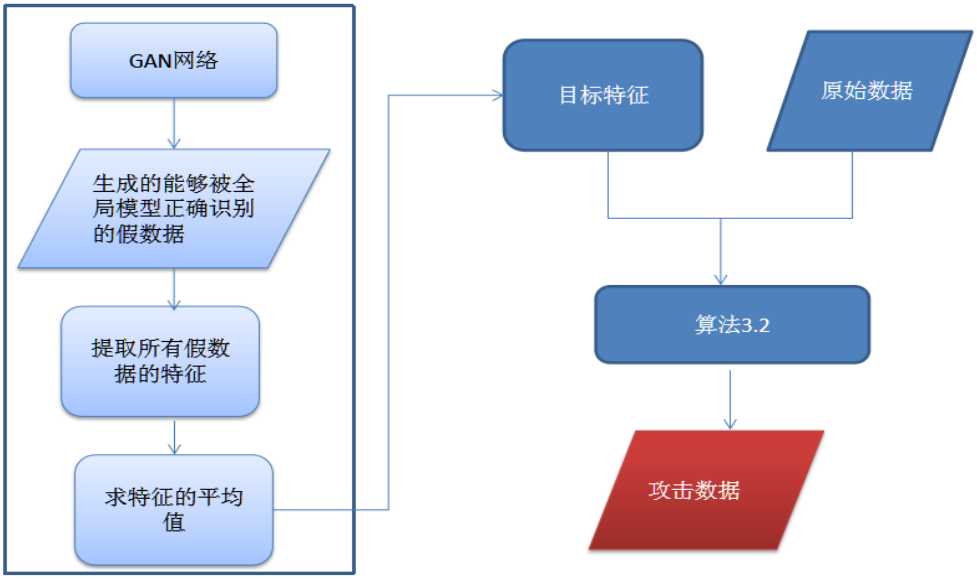


图 4.2 使用 GAN 网络生成的数据生成攻击数据

第二种是目标数据相近法。通过 GAN 网络生成指定标签的假数据，以生成的假数据为目标数据，在目标数据中选择与原始数据最相近的数据做为使用的目标数据。在目标数据中选择选择与原始数据最相近的数据有三种方法，分别是数据的特征差值的 2 范数最小，数据的原始值的差值的 2 范数最小，数据的特征差值的 2 范数和数据的原始值的差值的 2 范数之和最小。选择到最相近的目标数据后，以目标数据的特征使用基于特征的攻击数据生成算法处理原始数据最终生成攻击数据。

4.4 实验

实验环境：阿里云桌面、Inter(R) Xeon(R)KVM CPU @2.50GHz 2.50GHz 处理器，8GB 内存，64 位 Windows 操作系统，Python 程序语言，开发框架 PyTorch，

数据集 CIFAR-10,模型 ResNet18。联邦学习模型使用典型的服务器-客户端模式，10 个参与者其中 1 个为恶意参与者，1 个模型参数服务器。

4.4.1 使用生成对抗网络生成指定类别数据

实验中使用的联邦学习全局模型由10个参与者训练50轮得到，其中一个为恶意参与者，但是恶意参与者不做恶意行为，只是正常参与联邦学习即可。通过训练得到一个识别率为87.46%的模型。后续使用此模型的参数来更新生成对抗网络中的判别器D。

在实验中选择了数据集CIFAR-10中的所有标签，对每种指定标签分别训练50轮，每轮20次，每轮训练完成后输出生成的假图片。然后把生成的假图片通过全局模型选出可以被全局模型正确识别的图片作为目标图片。最后在目标图片中每个类别选取50张图片作为最终使用的目标图片。生成图片如图4.3所示，每个类别取了40张图片合成为一张图片。这些图片和对应标签的原始图片从外观上很难看出相似性，但是他们的确可以被联邦学习的全局模型正确识别为指定标签，这也反映出智能模型的脆弱性。如果模型没有一定的安全保护或者数据检测机制，这些生成的假数据就可以做到100%的攻击。在本方案中我们需要的是能够被模型正确识别为指定标签的图片而不需要在外观上去趋向真实图片，只要这些图片能够被模型正确识别，就可以做为目标图片用来处理原始图片从而生成攻击图片。

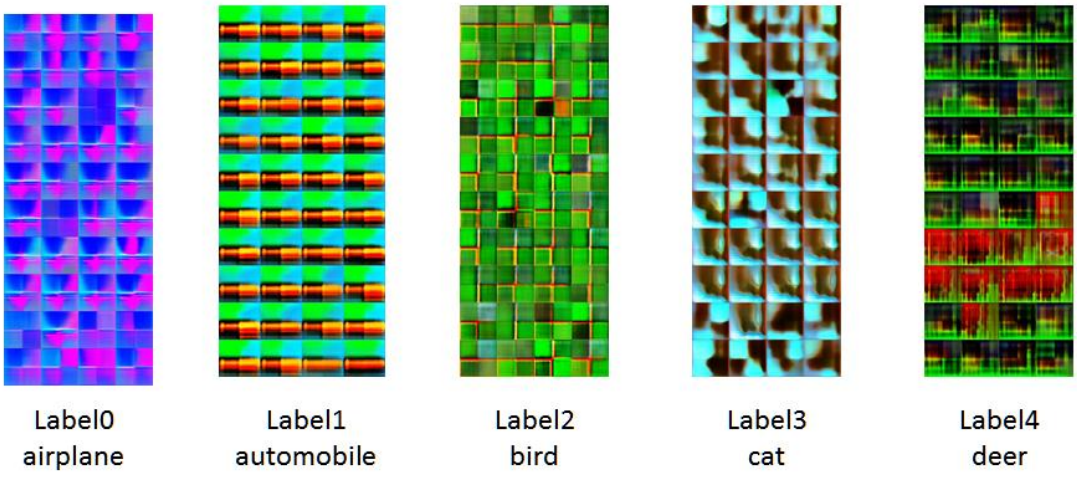


图 4.3 使用 GAN 网络生成的不同标签的假数据

4.4.2 使用生成的假数据生成攻击数据

第一种特征值平均法。在实验中使用目标特征值平均法来生成攻击数据，通过 500 轮的训练来生成攻击图片，每种类别生成的目标数据为 50 张图片，提取每一张图片的的特征，然后对这些特征求平均值，得到最终使用的目标特征。以得到的目标特征为目标使用算法 3.2 来处理原始图片最后生成攻击图片，每一个标签生成 50 张攻击图片，最终的攻击成功率如图 4.4 所示。从图中可以看出攻击率

平均在 40%以上，但是有的特别高比如类别 6，有的比较低比如类别 5 和类别 7，后续还需要继续改定生成算法，提高整体的攻击成功率。从图中可以看出只要恶意参与者参加了训练，得到了全局模型，恶意参与者就可以任意的攻击模型，而且成功率平均 40%以上，这种危害还是很大的。所以我们要不断地加强联邦学习本身的安全，构建更加安全的联邦学习。

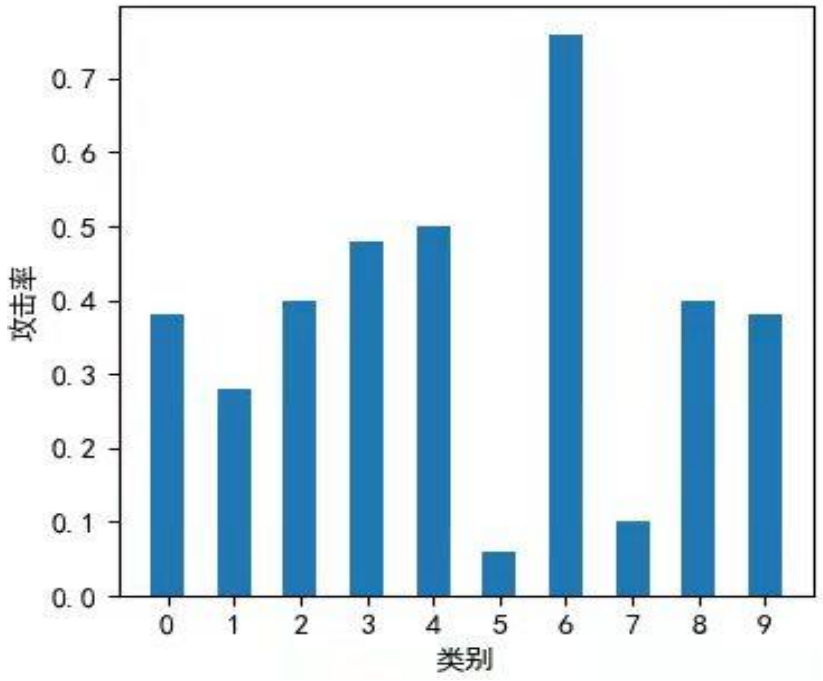


图 4.4 基于 GAN 网络和特征的后门攻击成功率

第二种目标数据相近法。恶意参与者使用 GAN 网络生成指定标签的数据，每种标签生成 50 张可以被模型正确识别的图片做为目标图片。以攻击鸟类图片为例说明如何生成攻击图片。随机取 50 张非鸟类原始图片，对每一张原始图片在 50 张目标图片中选择一张最相近的作为目标图片进行处理，以算法 3.2 生成攻击图片。选择最相近的方式有三种：1.两张图片特征最相近 2.两张图片原始值最相近 3.两张图片的特征距离和原始值距离之和最相近。三种方式以 2 范数表示距离。对十种类别图片分别按照上述方法生成攻击图片。在采用算法 3.2 的情况下，三种相近法生成的攻击图片攻击成功率和目标特征值平均法生成的攻击图片的攻击率对比图如图 4.5 所示。从图中可以看出原始值相近的方法的效果是最好的，做为攻击者可以轮换使用这些方法最大化攻击效果。

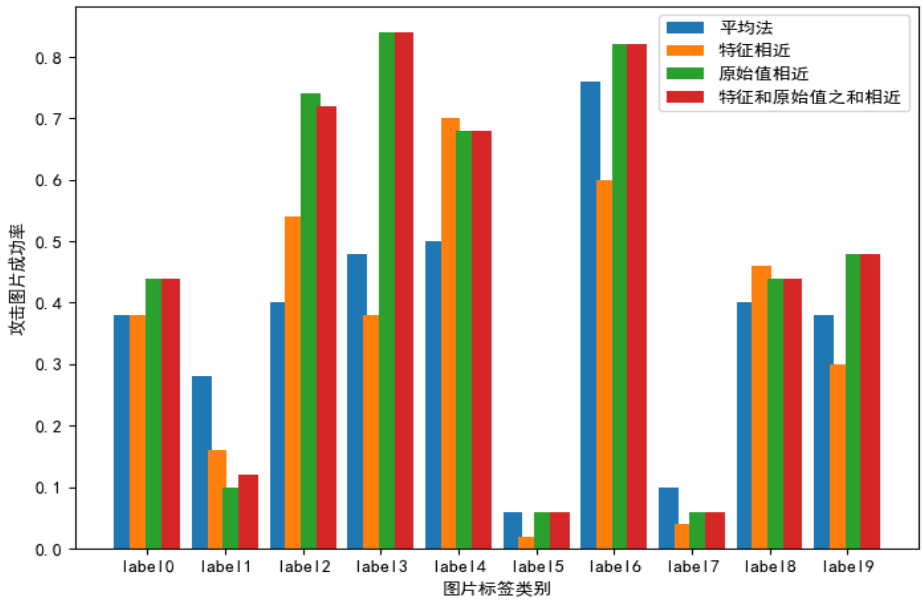


图 4.5 基于 GAN 网络生成数据的四种方法的攻击成功率

4.4.3 基于 GAN 网络的后门攻击方案对比分析

在后门攻击中分为两步，第一步在训练中植入后门，第二步在待检测数据中植入触发器触发后门。根据这两步我们对本文方案和文献[21]使用的方案进行对比如表 4.1 所示。通过对比我们可以发现本方案的最大优点在于不需要改变参与者的训练过程，可以让恶意参与者以一个正常参与者的状态参与训练，不做任何恶意行为，这样提高了恶意参与者参与联邦学习的概率。同时本文方案还可以实现对任意类别数据的攻击，而文献[21]的方案只能对植入了后门的类别实行攻击，灵活性远不及本文方案，最后在攻击成功率上本文方案并不占优，后续需要进一步改进。

表 4.1 基于 GAN 网络的两种后门攻击方案对比

方案名称	是否需要改变训练过程	如何改变训练过程	在训练中是否需要植入后门数据	是否需要改变待分类数据	攻击范围	攻击成功率
本文方案	否	否	否	需要利用本文所提方案处理待分类数据生成新的攻击数据	可以攻击任意类别数据	10 个类别平均攻击成功率为 40%，最高的是类别 6 蛙类 76%，最低的是类别 5 狗类 6%
文献[21]方案	是	参与者需要增加 GAN 网络来生成投毒数据参与训练，从而植入后门。	是，植入通过 GAN 网络和水印技术生成的投毒数据作为后门。	需要给待分类数据添加触发器。	可以攻击植入了后门的类别数据。	选用 GAN 网络生成的假数据中识别率最高的类别 2 Car,通过水印方式植入后门攻击成功率为 85%。

4.5 本章小结

本章详细介绍了基于生成对抗网络和特征的后门攻击方法的流程和实现过程，介绍了如何定义生成器 G，如何利用联邦学习全局模型更新判别器 D，以及如何利用生成器 G 和判别器 D 来生成指定标签的假数据。介绍了生成假数据后，如何使用基于特征的后门攻击方法生成攻击数据，最后展示了攻击效果。

总结与展望

联邦学习的特点就是各参与者在不分享自己本地数据的情况下进行联合训练更好的模型，从而得到了更好的模型并保护了数据和隐私安全。在《数据安全法》等旨在保护数据和隐私安全的法律法规不断出台的大背景下，联邦学习日益受到重视和广泛应用。由于联邦学习的特性，联邦学习成为了机器学习领域的安全范式。但是联邦学习的特点也容易在应用过程中遭到攻击，研究攻击方式可以促进对于联邦学习本身安全的了解和重视，了解了联邦学习的安全缺陷就可以构建更加安全的联邦学习。

本文在研究联邦学习的后门攻击中主要是提出了两种联邦学习的后门攻击方法，一种是联邦学习下基于特征的后门攻击方法，一种是联邦学习下基于生成对抗网络和特征的后门攻击方法。两种方法的核心都是通过特征提取，使用前向后向分裂迭代法，使得生成的攻击数据在特征上趋向于目标数据，在外观上保持原始数据基本不变。模型在对生成的攻击数据进行分类时会很大可能性的把攻击数据分类为目标数据的类别，从而达到攻击模型的目的。本研究的贡献就是：

（1）在原有的研究中，在使用前向后向分裂迭代法时使用特征差值的 2 范数作为误差然后进行梯度下降。本研究中修改了误差的表达式，改用特征差值的向量作为误差然后进行梯度下降。在 RestNet18 模型中图片经过模型之后的特征就是一个长度为 512 的特征向量，然后使用两张图片的特征向量的差值作为误差。

（2）在联邦学习中根据攻击者所处的不同场景，提出了对应的后门攻击方法。一种是以恶意参与者参与正常训练的数据作为后门。恶意参与者参与联邦学习的过程中拥有自己的本地数据，如果恶意参与者拥有想要攻击的目标类型的数据，就可以以自己拥有的数据作为后门，通过本文提出的基于特征的攻击方法生成攻击图片。以 CIFAR-10 数据集为例，恶意参与者拥有部分标签为 2 的鸟类数据，同时恶意参与者希望将任意的测试数据都分类为标签为 2 的鸟类数据。这时候恶意参与者在训练过程中不用植入后门只需正常的参与训练。等到需要攻击模型时，只需要把待分类的数据通过本文提出的方法让其在特征上趋向于恶意参与者拥有的鸟类数据特征，而在外观上基本不变。这种方法可以通俗的理解成就是把恶意参与者拥有的数据作为后门来使用，这种方式不用改变恶意参与者的训练过程，隐蔽性极高。第二种场景就是恶意参与者不拥有需要攻击的目标图片，需要植入后门。这种情形下，可以通过植入部分其他图片，标签设置为目标标签进行训练。需要攻击模型时，以植入的图片为目标来处理待分类图片从而生成攻击图片。在具体的实现模型攻击中，提出了目标特征值平均法和目标数据相近法两种方法，

然后通过实验对两种不同的方法的实验结果进行了比较。目标特征值平均法是在生成攻击图片时选用多张目标图片然后取平均值，最后以特征的平均值作为目标生成攻击图片。在目标数据相近法中分别讨论了三种不同的相近方法，分别是特征相近、原始值相近、原始值和特征值之和相近。在具体攻击中可以在这几种攻击方式中选择最优的攻击方式作为最终的攻击方式。

（3）提出了一种基于生成对抗网络和特征的联邦学习后门攻击方法。生成对抗网络以博弈的方式训练生成网络和判别网络，使得生成网络可以生成以假乱真的图片。联邦学习下，参与者可以得到全局的模型参数，就可以以全局模型作为判别器来训练生成器，生成器生成数据后无需以此植入后门，而是直接用此生成数据在基于特征的攻击方法下生成攻击数据，从而达到攻击模型的目的。

本研究还存在很多不足之处以及今后改进的方向:

(1)仅使用了 ResNet18 模型和 CIFAR-10 数据集，后续研究中可以使用不同的模型和数据集再进一步验证此方法的适用性和效率。

(2)图片通过 ResNet18 模型生成之后会生成一个长度为 512 的向量特征，这些特征在最后线性回归映射为对应的类别时的重要作用应该是不一样的，可以在后续的研究中加入注意力机制，寻找最重要的特征，然后以最重要的特征为目标来生成攻击数据，提高攻击的成功率。

参考文献

[1] 乐平公安. 2020 年全球数据泄露大事件盘点：数据“裸奔” 代价沉重. https://baijiahao.baidu.com/s?id=1690930815470412325&wfr=spider&for=pc, 2021.

[2] 澎湃新闻. 西宁市公安局召开“3.05”侵犯公民个人信息专案新闻发布会. https://m.thepaper.cn/baijiahao_13783323，2021.

[3] 百度百科. 通用数据保护条例. https://baike.baidu.com/item/通用数据保护条例/22616576?fr=aladdin.

[4] 百度百科. 中华人民共和国网络安全法. https://baike.baidu.com/item/中华人民共和国网络安全法/16843044?fr=aladdin.

[5] 百度百科. 中华人民共和国数据安全法. https://baike.baidu.com/item/中华人民共和国数据安全法/22861124?fr=aladdin.

[6] McMahan Brendan H，Moore Eider，Ramage Daniel，Hampson Seth，andBlaise Aguera Arcas．Communication-efficient learning of deep networks from decentralized data [C].In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics．Fort Lauderdale, FL, USA .2017: 1273–1282.

[7] 周俊，方国英，吴楠. 联邦学习安全与隐私保护研究综述 [J]. 西华大学学报(自然科学版)，2020，39(4): 9-17.

[8] 陈兵，成翔，张佳乐. 联邦学习安全与隐私保护综述［J］. 南京航空航天大学学报，2020，52（5）：675-684.

[9] 杨庚，王周生. 联邦学习中的隐私保护研究进展 [J]. 南京邮电大学学报 (自然科学版) , 2020, 40(5) : 204 -214.

[10] 周纯毅, 陈大卫, 王尚,等. 分布式深度学习隐私与安全攻击研究进展与挑战 [J]. 计算机研究与发展, 2021, 58(5):17.

[11] FREDRIKSON M，JHA S，RISTENPART T. Model inversion attacks that exploit confidence information and basic countermeasures [C].In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security．Denver, Colorado, USA . 2015: 1322 -1333.

[12] LI T, SAHU A K, TALWALKAR A, et al. Federated learning: challenges, methods, and future directions [J]. IEEE Signal Processing Magazine, 2020，37(3): 50 -60.

[13] MELIS L，SONG C Z，DE CRISTOFARO E，et al．Exploiting unintended

feature leakage in collaborative learning [C]. In: IEEE Symposium on Security and Privacy. San Francisco . 2019：691 -706.

[14] JAGIELSKI M，OPREA A，BIGGIO B，et al. Manipulating machine learning：Poisoning attacks and countermeasures for regression learning [C]. In: Proceedings of 2018 IEEE Symposium on Security and Privacy（SP）. San Francisco . 2018：19-35.

[15] XIAO H，BIGGIO B，NELSON B，et al. Support vector machines under adversarial label contamination [J]. Neurocomputing，2015，160：53-62.

[16] WANG B，YAO Y，SHAN S，et al. Neural cleanse：Identifying and mitigating backdoor attacks in neural networks [C]. In: Proceeding of 2019 IEEE Symposium on Security and Privacy（SP）. San Francisco . 2019：707-723.

[17] Bagdasaryan E, Veit A, Hua Y, et al. How to backdoor federated learning [C]. In: International Conference on Artificial Intelligence and Statistics. Naha, Okinawa, Japan. PMLR, 2020: 2938-2948.

[18] Sun Z, Kairouz P, Suresh A T, et al. Can you really backdoor federated learning? [J]. arXiv preprint arXiv: 1911.07963, 2019.

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

[20] Zhang J, Chen B, Cheng X, et al. PoisonGAN: Generative Poisoning Attacks Against Federated Learning in Edge Computing Systems [J]. IEEE Internet of Things Journal, 2020, PP(99):1-1.

[21] 陈大卫, 付安民, 周纯毅,等. 基于生成式对抗网络的联邦学习后门攻击方案[J]. 计算机研究与发展, 2021, 58(11):10.

[22] Bhagoji A N, Chakraborty S, Mittal P, etal. Analyzing federated learning through an adversarial lens [C]. In: International Conference on Machine Learning. Long Beach .PMLR, 2019: 634-643.

[23] Gu T, Dolan-Gavitt B, Garg S. Badnets: Identifying vulnerabilities in the machine learning model supply chain [J]. ArXiv preprint arXiv:1708.06733, 2017.

[24] WANG Z, SONG M, ZHANG Z, et al.Beyond inferring class representatives:User-level privacy leakage from federated learning [C]. In: Proceedings of IEEE INFOCOM Conference on Computer Communications. Paris.2019:2512-2520.

[25] SHEN S, TOPLE S,SAXENA P. Auror：Defending against poisoning attacks in collaborative deep learning systems [C]. In: Proceedings of the 32nd Annual

Conference on Computer Security Applications.USA. IEEE,2016: 508-519.

[26] CAO D, CHANG S, LIN Z, et al.Understanding distributed poisoning attack in federated learning [C]. In: Proceeding of the 25th International Conference on Parallel and Distributed Systems(ICPADS). China. IEEE，2019：233-239.

[27] ABADI M, CHU A, GOODFELLOW I, et al. Deep learning with differential privacy [C]. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. Austria. ACM，2016：308-318.

[28] Lapets A, Volgushev N, Bestavros A, et al. Secure MPC for analytics as a web application [C]. In: 2016 IEEE Cybersecurity Development (SecDev). Boston . IEEE, 2016: 73-74.

[29] Edo Roth, Daniel Noble, Brett Hemenway Falk, and Andreas Haeberlen．Honeycrisp: large-scale differentially private aggregation without a trusted core [C]. In: In Symposium on Operating Systems Principles. Huntsville, Ontario, Canada . 2019：196－210.

[30] David Lie and Petros Maniatis．Glimmers: Resolving the privacy/trust quagmire [C]. In: In Proceedings of the 16th Workshop on Hot Topics in Operating Systems. Whistler, BC, Canada. 2017：94－99.

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

[32] BONAWITZ K，IVANOV V，KREUTER B，et al．Practical secure aggregation for federated learning on user-held data [C].In: Proceedings of NIPS. Spain. MIT Press，2016：1015-1019

[33] SATTLER F，WIEDEMANN S，MÜLLE K R，et al．Robust and communication-efficient federated learning from non-iid data [J]. IEEE Transactions on Neural Networks and Learning Systems，2019，31(9)：3400-3413.

[34] Anish Athalye，Nicholas Carlini，and David Wagner．Obfuscated gradients give a false sense of security：Circumventing defenses to adversarial examples [C]. In: International Conference on Machine Learning .Sweden. ICML, 2018：274-283.

[35] Shafahi A, Huang W R, Najibi M, et al. Poison frogs! targeted clean-label poisoning attacks on neural networks [J]. Advances in neural information processing systems, 2018, 31.

[36] Yang Q, Liu Y, Chen T, et al. Federated Machine Learning: Concept and Appli-

cations [J]. ACM Transactions on Intelligent Systems and Technology, 2019, 10(2):1-19.

[37] Bonawitz K, Ivanov V, Kreuter B, et al. Practical secure aggregation for federated learning on user-held data [J]. arXiv preprint arXiv:1611.04482, 2016.

[38] McMahan B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data [C]. In: Artificial intelligence and statistics. Fort Lauderdale, FL, USA . PMLR, 2017: 1273-1282.

[39] Su H, Chen H. Experiments on parallel training of deep neural network using model averaging [J]. ArXiv preprint arXiv:1507.01239, 2015.

[40] Yu C, Tang H, Renggli C, et al. Distributed learning over unreliable networks [C]. In: International Conference on Machine Learning. Long Beach. PMLR, 2019: 7202-7212.

[41] Yu H, Yang S, Zhu S. Parallel restarted SGD with faster convergence and less communication: Demystifying why model averaging works for deep learning [C]. In: Proceedings of the AAAI Conference on Artificial Intelligence. Hawaii .2019, 33(01): 5693-5700.

[42] Liang G, Chawathe S. Privacy-preserving inter-database operations [C]. In: International Conference on Intelligence and Security Informatics. Springer, Berlin, Heidelberg, 2004: 66-82.

[43] Scannapieco M, Figotin I, Bertino E, et al. Privacy preserving schema and data matching [C]. In: Proceedings of the 2007 ACM SIGMOD international conference on Management of data. New York . 2007: 653-664.

[44] Rivest R L, Adleman L, Dertouzos M L. On data banks and privacy homomorphisms [J]. Foundations of secure computation, 1978, 4(11): 169-180.

[45] Alperin-Sheriff J, Peikert C. Faster bootstrapping with polynomial error[C]. In: Annual Cryptology Conference. Springer, Berlin, Heidelberg, 2014: 297-314.

[46] Brakerski Z, Vaikuntanathan V. Efficient fully homomorphic encryption from (standard) LWE [J]. SIAM Journal on computing, 2014, 43(2): 831-871.

[47] Ducas L, Micciancio D. FHEW: bootstrapping homomorphic encryption in less than a second [C]. In: Annual International Conference on the Theory and Applications of Cryptographic Techniques. Springer, Berlin, Heidelberg, 2015: 617-640.

[48] López-Alt A, Tromer E, Vaikuntanathan V. On-the-fly multiparty computation on the cloud via multikey fully homomorphic encryption [C]. In: Proceedings of the forty-fourth annual ACM symposium on Theory of computing. Palo Alto.

2012: 1219-1234.

[49] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition [C]. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas. 2016: 770-778.

[50] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets [J]. Advances in neural information processing systems, 2014, 27.

[51] Goldstein T, Studer C, Baraniuk R. A field guide to forward-backward splitting with a FASTA implementation [J]. ArXiv preprint arXiv:1411.3406, 2014.

致　谢

35 岁时开启了读研之路，此后三年便是人生最辛苦的几年，但是咬牙坚持了下来，今年 38 岁终于要毕业了。一路走来，虽然辛苦，但是也有学有所得的充实感和快乐感，在辛苦过后回首这几年，这几年的辛苦反倒成了人生中一段难忘的经历沉淀到了骨子里。在这三年中，和一群可以叫自己叔叔的小伙伴们一起学习，一起科研，真的是一件幸福快乐的事情。虽然自己年龄大了些，学习能力不如以前，而且家庭和工作上也有许多事情需要处理，但还是一路坚持走了下来，感谢自己没有放弃自己。在这三年读研路上要感谢的人着实很多。首先要感谢的就是自己的导师曹来成教授，曹老师当年在选择学生的时候并没有因为我年龄大而放弃我，而是一视同仁的接纳了我，给了我一次当曹门学子的机会。在这三年中，老师无时无刻不在关心着我，在学习上老师提前为我规划学习方向，在科研上老师悉心教导，手把手教我做科研，从选题的指导到开题报告的一字一句的修改，从做实验的安排到写论文的严格要求，无时无刻不体现了老师作为师者的担当和对学生的关爱。自己也是一名高职院校的老师，从曹老师身上学到了一个师者所应该有的品质，对学生在学习上科研上要关爱指导，在生活中上要关心关爱，在要求上要严格，对工作要认认真真兢兢业业。人生中遇到一名好老师是非常幸运的，无疑自己是幸运的，非常感谢曹老师三年的辛苦付出。

感谢学院的其他老师，给我传授了知识，树立了人生榜样。感谢辅导员金老师、吴老师这几年的辛苦付出，谢谢，祝老师们工作顺利，身体健康。

感谢师兄康一帆、吴琪瑞，师姐王玮婷、吴蓉在学习科研上的指导和帮助。还记的刚入学校时，师兄师姐们带我们认识校园，帮我们选课，经常组织活动让大家相互了解增进友情，在选题和写论文时悉心指导，非常感谢。感谢 19 级的小伙伴们李运涛、张文涛、张斐、王畋懿的帮助和鼓励，自己年龄大，这些小伙伴们还是很尊老的。从这些小伙伴们身上学到了很多东西，他们的年轻活力，他们的认真勤奋，他们的思维活跃，他们的善良可爱，都让我记忆深刻，感谢有你们。感谢师弟崔佐凯、吴文涛、李安、师妹何艳宁、朱敏对我的帮助和关心，也祝这些小可爱们学业有成。感谢舍友陶冶、李小强，还有形影不离的罗富元，我们经常一起学习，一起讨论，一起吃饭，一起跑步，一起玩，不是亲人胜似亲人，祝他们前程似锦，走向人生巅峰。

感谢单位的领导和同事对我的帮助，让我可以完成学业，感谢你们在我最困难的时候给与我的帮助，谢谢。

感谢家人一直对我的支持，这几年我在上学，爱人也在上学，孩子都是由父

亲母亲帮忙照顾，感谢父母这个坚强的后盾，让我可以安心读完这三年，完成这个学业。感谢乖巧的儿子，每次离别时总是来一句爸爸加油，让我有了无尽动力，而且儿子身体还不错，很少生病，让我少操好多心，儿子给力。最后还要感谢爱人，我们一起上学，虽不在同一个地方，但是可以相互交流、相互鼓励、相互自嘲，忘不了最困难的时候两个人相互自嘲是两个学渣。

宝剑锋从磨砺出，腊梅香自苦寒来，经历了三年的磨练，学到了知识，收获了友情，感谢这些美好时光。

附录 A　攻读学位期间所发表的学术论文目录

[1] Laicheng Cao, Fengqiang Li. Backdoor attack based on feature in federated learning [C]. In: 2022 International Conference on Internet and Cyber Security Technology. Guilin-China . 已接受，待发表，文章编号 IVAQYIPZXU.（对应学位论文第三章）

附录 B 攻读硕士学位期间参与的科研项目

[1] 国家自然科学基金项目，面向多用户动态可搜索隐私保护的云存储服务机制(61562059).