# 联邦学习隐私保护机制综述

摘要：

Abstract：

## 引言

近年来，机器学习算法在人工智能领域内迅猛发展，在计算机视觉、自然语言处理和推荐算法等领域都有良好的表现。然而，机器学习算法都是以大量数据为驱动的，这些数据可能分布在不同的设备或者组织上，并且含有敏感信息，因此，在数据隐私保护的限制下，直接将这些数据聚合在一起是不可行的。

为了解决上述数据孤岛问题，一种可行的方案是谷歌的H.Brendan McMahan等人于2016年提出的联邦学习（《Communication-Efficient Learning of Deep Networks from Decentralized Data 》）（Federated Learning）。联邦学习是多个实体（客户端）协作解决机器学习问题的一种方案，它在一个中央服务器或服务提供商的协调下进行，每个客户端的原始数据存储在本地，无法交换或迁移，满足了隐私保护和数据安全的要求。谷歌的研发人员开发了一个联邦学习系统，通过智能手机的私人数据来更新其Gboard系统（一种虚拟键盘系统）的输入预测模型，通过联邦学习技术，所有的智能手机的数据都可以被用来优化模型，而这一过程并不需要将数据从智能手机发送到中央设备。通过联邦学习技术，每台移动设备对自身训练得到的模型进行加密并上传到云端，所有的加密模型都被聚合到一个全局加密模型中，云服务器并不能得到每台移动设备的数据或者模型。聚合得到的全局模型会被广播给所有移动设备，然后移动设备进行解密，得到新的模型，完成模型更新的过程。在上述过程中，每台移动设备的个人数据都没有被透露给云端，因此是满足数据隐私保护的。

本文主要就联邦学习中的隐私保护机制进行了调查和研究，从隐私保护技术的角度对当前的研究联邦学习中的隐私保护的文献进行了分类，主要采用隐私保护技术有差分隐私、同态加密和安全多方计算，部分文献也混合使用了上述隐私保护技术的两种或三种，也有部分文献使用了其它隐私保护技术。

本文的组织结构如下：第2节简单介绍了有关差分隐私、同态加密和安全多方计算的概念和知识，并且回顾了联邦学习的威胁模型；第3节对现有联邦学习中隐私保护技术的研究进行了分类并且介绍了隐私保护技术的基本原理和研究现状，将主要从差分隐私、同态加密、安全多方计算、混合模式和其它方法，一共五个方面进行介绍。第4节将总结本文的工作和未来联邦学习隐私保护问题的研究方向。

## 背景知识

本节将介绍联邦学习中常用的隐私保护技术和联邦学习的威胁模型。联邦学习中需要对用户上传到云端的梯度或者模型进行保护，因为模型和梯度都会泄露用户的隐私数据，常用的隐私保护的思想是对梯度进行加密或者掩饰，前者可以使用同态加密，后者可以是差分隐私或者安全多方计算。

### 2.1 威胁模型

联邦学习不可能总是提供足够的隐私保证，也会遭受潜在的隐私攻击，危及模型和数据的完整性。文献（《Threats to Federated Learning：A Survey》）详细的介绍和总结了联邦学习面对的威胁和攻击。

在联邦学习的威胁模型中，通常会受到来自内部或外部的攻击，其中内部攻击往往比外部攻击要强。对于联邦学习的内部攻击，可以采取单一攻击，拜占庭攻击和女巫攻击。

根据主动性可以将对手分为半诚实的对手和恶意的对手，半诚实的对手被认为是诚实但好奇的，在不违背联邦学习协议的情况下试图了解其他方的私密状态；而一个活跃或恶意的对手可以通过修改、重放或删除消息来随意违背联邦学习协议，还可以进行毁灭性的攻击。

联邦学习可以分为训练和推理两个阶段，所以每个阶段受到的攻击也不同。训练阶段的攻击企图学习、影响和破坏模型本身，对手可以用数据投毒攻击和模型投毒攻击，还可以对参与者的更新发动一系列推理攻击。推理阶段的攻击通常不会篡改目标模型，只会导致模型产生错误的输出，或收集关于模型特征的证据。

投毒攻击主要有本地数据采集时的数据投毒攻击和在局部模型训练过程中的模型投毒攻击。数据投毒可以分为clean-label和dirty-label攻击，clean-label攻击假设对手无法改变任何训练数据的标签；相比dirty-label攻击中，对手可以将一些它希望用所需目标标签误分类的数据样本引入到训练集中。dirty-label投毒攻击的常见方式是标签翻转攻击和后面攻击。模型投毒攻击的目标是在将本地模型更新发送到服务器之前进行投毒，或者在全局模型中插入隐藏的后门。

推理攻击则可以分为成员推理攻击和属性推理攻击。成员推理攻击的目的是确定某一数据点是否被用于训练模型，而属性推理攻击的用来推断其他参与者的训练数据的属性。

### 2.2 差分隐私

差分隐私（《Calibrating noise to sensitivity in private data analysis》）是由Dwork在2006年首次提出的，它提供了量化和限制个人信息泄露的一种输出隐私保护模型。差分隐私的中心思想是，当攻击者试图从数据集中查询个体信息时将其混淆，使得敌手无法从查询结果中辨别个体的敏感性，即函数的输出结果对于数据集中的任何特定记录都不敏感，因此，差分隐私能被用来抵抗成员推理攻击。差分隐私定义如下：

差分隐私：一个随机化机制，其定义域为，如果满足差分隐私，那么对于所有的和满足的，有：

式中，表示隐私预算；表示失败概率。一般而言，越小，隐私保护程度越高，噪声越大，数据可用性越差。

差分隐私的主要实现方式是向数据添加噪声，主要有两种方式，一种是根据数值型的输出函数的敏感度添加噪声，比如基于敏感度的拉普拉斯噪声 （Laplace Noise）和敏感度高斯噪声（Gaussian Noise）；另一种是根据离散值的指数分布选择噪声。

### 2.3 同态加密

### 2.4 安全多方计算

安全多方计算（Secure Multi-Party Computaion，SMPC）最早是由图灵奖获得者、中国科学院院士姚期智于1982年正式提出的（《Protocols for secure computations》）， 其目的是从每一方的隐私输入中协作计算一个函数的结果，而不用将这些输入展示给其它方。安全多方计算保证了参与方获得正确结果的同时，无法获得计算结果之外的任何信息。

安全多方计算可以通过三种不同的框架来实现：不经意传输（Oblivious Transfer，OT）、秘密共享（Secret Sharing，SS）和阈值同态加密。本文将在下面内容中简单介绍不经意传输和秘密共享。

不经意传输是一种由Rabin在1918年提出的两方安全协议（《How to exchange secrets with oblivious transfer》），在不经意传输中，发送方拥有一个“索引-消息”对，每次传输时，接收方选择一个索引，并接受对应的消息。接收方不能得知数据库的任何其它消息，发送方也不了解接受方选择的索引的任何信息。

秘密共享（Secret Sharing）是指将秘密信息分割成若干份，然后将这些秘密份额分给不同的人保管，以达到隐藏秘密和风险分散的目的。一般来说，一个秘密分享方案包含了一个秘密分割算法和一个秘密重建算法，包含了分发者、持有者和接受者三种角色。分发者负责将秘密信息通过秘密分割算法进行分割，并发送给持有者。接收者在需要秘密的时候对秘密进行重建，收集一组持有者的秘密份额，并执行秘密重建算法来恢复秘密，当有足够的秘密份额时就可以得到秘密信息。常用的一种秘密共享的方法是Shamir秘密共享（Shamir’s Secret Sharing），是基于多项式方程构建的。

## 隐私保护机制

联邦学习中目前主要使用的隐私保护机制有三种：差分隐私、同态加密和安全多方计算，部分研究人员也混合使用了两种或者三种来实现更好的隐私保护，部分研究人员也采用了一些其它的隐私保护方式。本节将从这个五个方面来介绍联邦学习中的隐私保护机制。

### 3.1 差分隐私

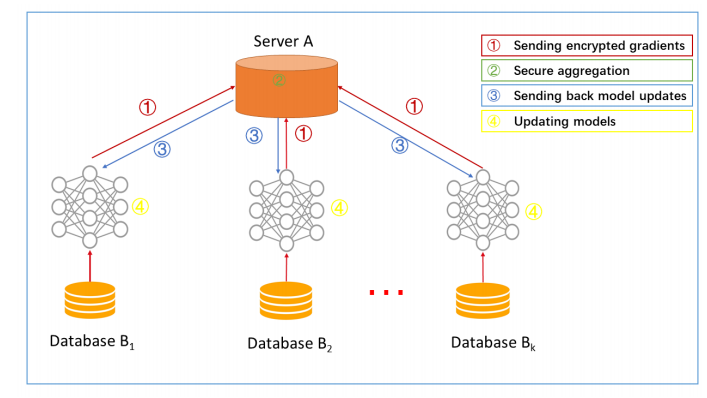
差分隐私机制通常被应用到联邦平均算法FedAvg（Communication-Efficient Learning of Deep Networks from Decentralized Data）中，算法的具体过程如图1所示，每个客户端在本地计算模型梯度，然后将梯度发送给中央服务器，服务器进行梯度聚合，使用聚合梯度进行模型更新，得到新的全局模型并广播给所有客户端。服务器直接将模型广播和发布可能会泄露成员信息，恶意攻击者可以通过成员推理攻击来推测某个客户端是否参与了训练；而客户端直接将梯度发送给服务器也会向服务器泄露隐私信息，服务器可以通过梯度来推断客户端的数据集中有无某个样本，差分隐私算法可以应用到联邦平均算法的过程中，通过添加噪声的方式以抵抗推理攻击。差分隐私算法可以根据噪声添加的位置分为用户级和样本级，样本级差分隐私算法也可以称为本地差分隐私（Local Different Privacy，LDP）。

图1：联邦平均算法

文献（《Learning Differentially Private Recurrent Language Models》）率先将差分隐私技术应用到联邦平均算法中。在本地客户端更新过程中，使用参数裁剪来限制样本梯度的大小，在聚合全局模型时，使用差分隐私的高斯机制对缩放后的模型更新添加噪声，使用Moment accountant（《Deep Learning with Differential Privacy》）算法计算总体隐私损失，实现了用户级别的隐私保护。文献的实验表明，在一个具有足够多用户的数据集上，实现差分隐私是以增加计算量为代价的，在大型数据集上进行训练时，私有LSTM语言模型在数量和质量上都与无噪声模型相似。文献（《Differentially Private Federated Learning :A Client Level Perspective》）也使用了类似的方法在客户端级别使用差分隐私来隐藏客户端的贡献，有所不同的是，前者在本地客户端进行训练过程中进行参数裁剪，而后者在中央服务器进行参数裁剪，二者参数裁剪的阈值也不相同。文献还指出在联邦学习训练过程中动态调整差分隐私机制可以提高模型性能。文献通过实验表明，在有足够数量的参与客户的情况下，其程序可以在模型性能上以较小的成本维持客户级的差异隐私。

文献（《Differentially Private Learning with Adaptive Clipping》）改进了文献的参数裁剪过程，提出了一种新的自适应的分位数裁剪策略，用于训练具有用户级别的差分隐私的联邦学习模型，从而无需进行大量参数调整。文献指出为样本添加限制对于学习过程和差分隐私都是必要的，之前设置固定的限制的方法会导致学习模型实用性的降低。文献描述了自适应分位数裁剪策略，该策略是为迭代隐私机制设计的，然后描述了特定层的噪声添加策略，后者比基本策略具有更高的实用性。

文献（《cpSGD: Communication-efficient and differentially-private distributed SGD》）对二项机制进行了改进和扩展，应用到了DEM问题中（distributed mean estimation，DEM，分布式向量平均估计，在联邦学习中，分布式向量是模型梯度或参数），主要思想是：对于每个客户端，在给服务器发送梯度之前，将经过适当参数化的二项式分布得出的噪声添加到每个量化值，服务器进一步减去噪声引入的偏差，以实现无偏差的均值估计器。文献还进一步表明，随机轮换有助于减少因差分隐私导致的附加错误。

文献（《Protection Against Reconstruction and Its Applications in Private Federated Learning》）针对大规模分布式学习和联邦学习，提出了一种针对不同隐私要求的统计学习问题的最小最大（minimax）差分隐私机制，提出了一个新的联邦学习隐私保护框架：在客户端本地计算模型参数更新的过程中，使用自身提出的本地隐私保护机制保护本地数据，在中央服务器执行聚合过程中，使用差分隐私保证模型参数的通信过程是全局私有的。整个反馈环路提供了有效的隐私保护，用户的本地数据不会传输到中央服务器，而集中式隐私保护能够保证过程和最终参数都不存在敏感性披露的风险。

文献（《Differentially Private Meta-Learning》）研究了元学习背景下的隐私问题，提出了一种基于梯度元学习的差分隐私算法（DP GBML），该算法在每次任务内迭代时使用了高斯机制的差分隐私梯度下降，保护了任务中的单个样本的隐私，并在凸环境下具有学习保证。文献演示了如何将隐私概念转化为有用的深度学习模型，并证明了在联邦语言模型和小样本图像分类任务中具有出色的性能。

文献（《Federated Learning with Bayesian Differential Privacy》）提出了一种新的贝叶斯差分隐私框架和一种实用的隐私损失计算方法，贝叶斯差分隐私是对类似分布数据的差分隐私的放松版本。文献将差分隐私框架应用到联邦环境中，并使用提出的隐私损失计算方法估算客户的隐私保证；文献强调了样本级差分隐私的重要性，并提出了估计样本级隐私损失的两种变体；最后，文献介绍了一种新的联合计算隐私损失的方法，可以同时估计客户端级别和样本级别的隐私。在客户端具有相似的数据分布的场景下，文献的实验证明其方法可以减少隐私预算并且提高训练模型的性能。

文献（《FedSel Federated SGD under Local DifferentialPrivacy with Top-k Dimension Selection》）针对联邦梯度下降算法提出了一个两阶段的本地差分隐私框架FedSel，其关键思想是首次尝试通过延迟不重要的梯度减轻了注入噪声的维度问题。FedSel有两个阶段组成，分别是“维度选择”和“值扰动”，文章列举了三种“维度选择”的方法，并提供了对应的隐私保证。文献从理论上分析了FedSel的隐私、准确度和时间复杂度，其性能优于最新的解决方案，在现实世界和合成数据集上进行实验测试了FedSel框架的有效性和效率。

文章（《LDP-Fed Federated Learning with Local Differential Privacy》）提出了LDP-Fed，使用本地差分隐私（LDP）提供正式的隐私保证。LDP-Fed主要有两个新的组件：在每个客户端运行的本地差分隐私模块和k-客户端选择模块，前者使用指数机制和CLDP（本地差分隐私的变体）提供隐私保护，后者用于在迭代LDP-Fed训练过程的各个回合中选择性共享模型参数更新。文章根据最新的隐私保护FL方法评估了LDP-Fed的准确性和系统功能。

文献（《Exploring Private Federated Learning with Laplacian Smoothing》）将基于拉普拉斯平滑（Laplaian Smoothing）的训练效果增强方案应用到差分隐私联邦学习过程中（DP-Fed-LS），在梯度聚合时注入了高斯噪声，在模型精度上得到了改善。另外，文献为均匀二次抽样和泊松二次抽样提供了严格的封闭形式的隐私边界，得出了差分私有联邦学习的差分隐私保证，无论有无拉普拉斯平滑。文献通过一系列实验表明了，在不同的差分隐私保证和二次采样机制下，DP-Fed-LS提供的效用比DP-Fed更好。

文献（《Distributed Differentially Private Averaging with Improved Utility and Robustness to Malicious Parties》）提出了一种新颖的差分私有的平均协议——GOPA（Gossip Noise for Private Averaging）。该协议可以匹配可信管理者设置的准确性，同时自然的扩展到大量用户。客户端需要将两种噪声加到了私有值上，第一种噪声是客户端与相邻客户端交换的高斯噪声，该噪声在聚合时可以相互抵消；第二种噪声是独立的服从高斯分布的噪声，无法消除，是对一种噪声的补充，文献分析了GOPA协议的差分隐私保证和在面对恶意行为时的正确性。另外，文献通过承诺方案和零知识证明，保证了用户能够证明他们计算的正确性，而又不影响协议的效率和保密性。

### 3.2 同态加密

### 3.3 安全多方计算

### 3.4 混合模式

### 3.5 其它

## 总结

## 参考文献