**摘要**

基于人工神经网络的深度学习是非常流行的对图像，语音和文本等复杂数据进行建模，分类和识别的方法。深度学习方法前所未有的准确性已经使它们成为互联网上基于AI的服务的新的基础。大规模收集用户数据的商业公司是这种趋势的主要受益者，因为深度学习技术的成功是直接的与可用于训练的数据量成有关。

深度学习所需的大量数据收集存在明显的隐私问题。用户的个人的、高度敏感的数据，如照片和录音，被这些公司无限期地保存。用户既不能删除它，也不能限制它的使用目的。此外，集中保存的数据受制于法律传票和司法外监督。许多数据所有者 - 例如，可能希望将深度学习方法应用于临床记录的医疗机构 – 因为隐私和机密性问题不能共享/使用数据从而受益于大规模深度学习。

在本文中，我们设计，实现和评估一个实际系统，使多方能够共同学习给定目标的精确神经网络模型，而无需共享他们的输入数据集。我们利用现代深度学习中使用的优化算法，即基于随机梯度下降的优化算法，可以并行化并异步执行。我们的系统在训练中允许参与者在他们自己的数据集上独立训练并有选择地分享其模型关键参数的子集。这为实用/隐私权衡空间提供了一个吸引人的点：参与者保护其各自的隐私数据，同时仍然受益于其他参与者的模型将学习准确性提高到高于仅依靠自己的自己的输入可实现的水平。我们展示了我们在基准数据集上保护隐私的深度学习的准确性。

**关键词** 隐私; 神经网络; 深度学习; 梯度下降

**1简介**

基于人工神经网络的深度学习方法的最新进展已导致由来已久的人工智能任务的突破，如语音，图像和文本识别，语言翻译等。谷歌，Facebook和苹果等公司利用从用户收集的大量训练数据以及GPU场的巨大计算能力部署了大规模深度学习。所得模型的前所未有的准确性使它们成为许多新服务和应用的基础，包括准确的语音识别[24]和优于人类的图像识别[26]。

虽然深度学习的实用性是不可否认的，但使其如此成功的训练数据也会带来严重的隐私问题。数百万人的照片，演讲和视频的集中收集充满隐私风险。首先，收集这些数据的公司会永远保留数据，被收集数据的用户既不能删除它，也不能控制它的使用方式，也不会控制从中学到的东西。其次，图像和录音通常包含意外捕获的敏感物品 - 面孔，车牌，电脑屏幕，其他人说话的声音和环境噪音[44]等。第三，公司保存的用户数据受传票和保证的约束，以及国家安全和情报机构的无证间谍活动。

此外，互联网巨头对数百万用户收集的“大数据”的垄断导致他们垄断从这些数据中学到的AI模型。用户可以从新服务中受益，例如强大的图像搜索，声控个人助理以及外语网页的机器翻译，但是根据他们的集体数据构建的基础模型仍然是创建它们的公司的专有模型。

最后，在众多领域，最值得关注的是医学，法律或法规不允许共享有关个人的数据。因此，生物医学临床研究人员只能对属于他们自己的机构的数据集进行深度学习。众所周知，训练数据集变得越大，越多样化，神经网络模型就越好。由于在训练模型时无法使用其他机构的数据，研究人员最终可能会得到糟糕的模型。例如，由单个组织（例如，特定医疗诊所）拥有的数据可能是非常同质的，产生过度拟合的模型，当在其他输入上使用时该模型将是不准确的。在这种情况下，隐私和保密限制显着降低了效用。

***我们的贡献。***我们设计，实施和评估实用的协作深度学习系统，在效用和隐私之间提供了很有吸引力的权衡。我们的系统允许多个参与者在他们自己的输入上学习神经网络模型，而不共享这些输入，但从同时学习类似模型的其他参与者中受益。

我们的关键技术创新是在培训期间选择性地共享模型参数。该参数共享在随机梯度下降期间与局部参数更新交织，允许参与者在没有明确分享训练输入的情况下受益于其他参与者的模型。我们的方法独立于用于为特定任务构建模型的特定算法。因此，它可以轻松地适应未来神经网络训练的改进，而无需更改核心协议。

选择性参数共享是有效的，因为现代神经网络训练的随机梯度下降算法可以并行化并且异步运行。它们对于不可靠的参数更新，竞争条件，参与者退出等都是健壮的。用从其他参与者获得的值更新一小部分参数允许每个参与者在寻找最优参数的过程中避免局部最小值。可以调整参数共享以控制交换的信息量与所得模型的准确性之间的权衡。

我们在两个数据集MNIST和SVHN上实验评估我们的系统，这两者都用作图像分类算法的基准。我们系统中分布式参与者生成的模型的准确性接近于集中的，隐私违规的情况，其中单方持有整个数据集并使用它来训练模型。对于MNIST数据集，当参与者共享其参数的10％（分别为1％）时，我们获得99.14％的准确度（分别为98.71％）。相比之下，集中式隐私违规模型的最高准确率为99.17％，参与者个人学习的非协作模型的最高准确率为93.16％。对于SVHN数据集，当参与者共享其参数的10％（1％）时，我们达到93.12％（89.86％）的准确度。相比之下，集中式隐私违规模型的最高准确率为92.99％，非协作模型的最高准确率为81.82％。

即使没有额外的保护措施，我们的系统已经实现了比现有系统更强大的隐私，实用损失可以忽略不计。没有直接揭示所有训练数据的风险，唯一的我们系统中的泄漏是间接的，通过一小部分神经网络参数。为了尽量减少这种泄漏，我们将展示如何使用稀疏矢量技术将差异隐私应用于参数更新，从而减轻了由于参数选择（即，选择要共享哪些参数）和共享参数值而导致的隐私损失。然后我们定量测量准确性和隐私之间的权衡。

**2相关工作**

**2.1深度学习**

深度学习是从复杂数据学习非线性特征和函数的过程。深度学习架构，算法和应用程序的调查可以在[5,16]中找到。深度学习在语音识别[23,24,27]，图像识别[30,45]和面部检测[48]方面已被证明优于传统技术。基于新类型整流器激活函数的的深度学习架构在识别ImageNet数据集中的图像时[26]的性能据称优于人类。

深度学习有望分析与癌症[13,22,32]和遗传学[15,56] 相关的复杂生物医学数据。从隐私角度来看，用于构建这些模型的培训数据特别敏感，强调了保护隐私的深度学习方法的必要性。

我们的工作受到最近深度学习并行化进展的启发，特别是GPU / CPU集群上随机梯度下降的并行化 [14]，以及在神经网络训练期间分布式计算的技术[1,39,59]。但是，这些技术并不涉及训练数据的隐私，并且所有技术都假设单个实体控制训练过程。

**2.2机器学习中的隐私**

关于机器学习中隐私保护的现有文献主要针对传统的机器学习算法，而不是深度学习，并解决三个目标：用于学习模型或作为现有模型输入的数据的隐私，模型的隐私，以及模型输出的隐私。

基于安全多方计算（SMC）的技术可以在多方对其专有输入进行协作机器学习时帮助保护计算的中间步骤。SMC已被用于学习决策树[33]，线性回归函数[17]，关联规则[50]，朴素贝叶斯分类器[51]和k-均值聚类[28]。一般而言，SMC技术会产生非平凡的性能开销，并且它们在保护隐私的深度学习中的应用仍然是一个悬而未决的问题。

保护模型隐私的技术包括保护隐私的概率推理[38]，保护隐私的发言人识别[36]，还有计算加密数据[3,6,55]。通过相比之下，我们的目标是协同训练神经网络可以由每个参与者私下和独立使用。

差异隐私[19]是一种保护隐私的机器学习的流行方法。它已被应用于增强[21]，主成分分析[10]，线性和逻辑回归[8,57]，支持向量机[41]，风险最小化[9,53]和连续数据处理[43]。最近的研究结果表明，随机梯度下降的嘈杂变量实现了最小化Lipchitz凸函数在l2-有界集上的最优误差[4]，并且用于防止过度拟合的随机“dropout”也可以在一个简单的1层神经网络[29]中加强隐私保护。据我们所知，以前的工作都没有解决使用分布式随机梯度下降的多方协作深度学习的问题。

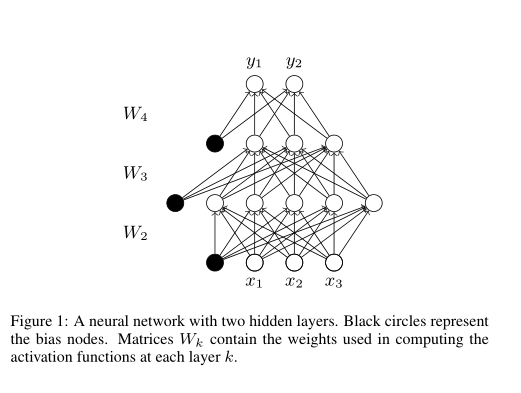
在[37]中提出了聚合使用差分隐私和安全多方计算独立训练的神经网络。不幸的是，平均神经网络参数确实如此不一定会产生更好的模型。

与先前提出的技术不同，我们的系统实现了协同神经网络训练背景下的全部三个隐私目标：它保护培训数据的隐私，使参与者控制学习目标以及揭示他们的个人模型，并让他们应用共同学习的模型到自己的输入上而不会泄露输入或输出。我们的系统以比加密技术（如安全的多方计算或同态加密）低得多的性能成本实现这一目标，并且适用于现代部署大规模的深度学习。

**3深度学习**

深度学习旨在从高维数据中提取复杂特征，并使用它们构建将输入与输出（例如，分类）相关联的模型。深度学习体系结构通常被构造为多层网络，因此更抽象的特征被计算为较低级特征的非线性函数。我们主要关注监督学习，其中训练输入被标记为正确的类，但原则上我们的方法也可以用于无监督的隐私保护的学习。

多层神经网络是最常见的深层形式学习架构。图1显示了典型两个隐含层的神经网络有。网络中的每个节点都模拟一个神经元。



在典型的多层网络中，每个神经元接收前一层中神经元的输出加上值为1的特殊神经元的偏置信号。然后计算其输入的加权平均值，称为总输入。此神经元的输出就是将总输入作为输入计算非线性激活函数。k层中神经元的输出向量是，其中f是激活函数，是权重矩阵，确定每个输入信号的贡献。激活函数的示例是

双曲正切，sigmoid ，整流器，softplus。如果神经网络用于将输入数据分类为有限数量的类（每个类由不同的输出神经元表示），最后一层的激活功能通常是

softmax函数，∀j。在这种情况下，每个的输出最后一层中的神经元j是输入属于j类的相对分数或概率。

通常，在较高层中计算的值表示数据更抽象特征。第一层由从数据中提取的原始特征组成，例如图像中每个像素的颜色强度或文档中每个单词的频率。最后一层的输出对应于模型产生的抽象答案。如果神经网络用于分类，这些抽象特征也表示输入和输出之间的关系。非线性函数f和权重矩阵确定在每一层提取的特征。深度学习的主要挑战是自动从训练数据中学习最大化神经网络目标的参数（权重矩阵）的值（例如，分类准确度）。

***使用梯度下降学习网络参数。***学习神经网络的参数是非线性优化问题。在监督学习中，*目标函数*是神经网络的输出。用于解决这个问题的算法通常是梯度下降的变种[2]。简单地说，梯度下降从随机点（参数集）开始然后，在每个步骤中，计算正在优化的非线性函数的梯度，并更新参数以减小梯度。这个过程一直持续到算法收敛于局部最优。

在神经网络中，每个权重参数的梯度是通过*前馈和反向传播*程序计算。前馈顺序计算给定输入数据的网络输出，然后计算误差，即该输出与功能的真实值之间的差值。反向传播通过网络传回此错误并计算每个神经元对总误差的贡献。根据神经元的激活值及其对误差的贡献计算各个参数的梯度。

***随机梯度下降（SGD）。***可以在所有可用数据上取参数梯度的平均值。这个算法，也称作*批量梯度下降*，效率不高，特别是如果学习上一个大数据集。随机梯度下降（SGD）是一种极大的简化，它可以在极小的整个数据集的子集（小批量）内计算梯度 [58]。在最简单的情况下，对应于最大随机性，在每个优化步骤中随机选择一个数据样本。

设是神经网络中所有参数的平坦向量，由， 组成。设E为*误差函数*，即目标函数的真实值与网络计算出的输出之间的差异。 E可以基于范数或交叉熵[34]。反向传播算法计算E的相对于w中的每个参数的偏导数并更新参数以减小其梯度。参数的随机梯度下降更新规则是

其中α是*学习率*，并且在小批量i上计算。我们将所有可用输入数据的一次完整迭代称为时期。

请注意，向量w中的每个参数都独立于其他参数进行更新。在本文的其余部分设计我们的隐私保护，协作随机梯度下降系统时，我们将依赖此属性。 一些技术自适应地设定学习率[18]，但仍保持这种独立性。

**4分布式选择性SGD**

我们的方法的核心是分布式协作深度学习协议，它依赖于以下观察：（i）在梯度下降期间对不同参数的更新本质是独立的，（ii）不同的训练数据集有助于不同的参数，以及（iii）不同的特征对目标函数的贡献不同。我们的选择性随机梯度下降（选择性SGD或SSGD）协议实现了与传统SGD相当的精度，但涉及在每次学习迭代中更新的参数少了1或甚至2个数量。

***选择性参数更新***。选选择性参数更新背后的主要直觉是在SGD期间，一些参数对神经网络的目标函数贡献更多，因此在给定的训练迭代期间经历更大的更新。梯度值取决于训练样本（小批量），并且因样本而异。此外，输入数据的某些特征比其他特征更重要，有助于计算这些特征的参数在学习过程中发挥更大的作用并经历更大的变化。

在选择性SGD中，学习者选择一小部分参数每次迭代都要更新。此选择可以是完全随机的，但聪明的策略是选择当前远离它们的局部最优值的参数值，即那些具有更大梯度值的参数。对于每个训练样本i，对于所有参数w j，计算偏导数，如SGD。设S是具有最大值的θ参数的序号。最后，以与（1）中相同的方式更新参数向量，因此不在S中的参数保持不变。我们将θ与参数总数之比称为参数选择率。

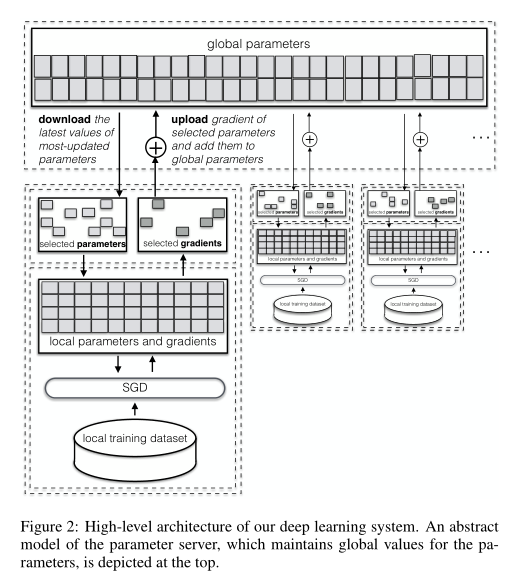
***分布式协作学习***。分布式选择性SGD假设两个或更多参与者独立和同时进行培训。在每轮本地训练之后，参与者彼此异步地共享他们为某些参数计算的**梯度**。每个参与者完全控制要分享的梯度和频率。针对给定参数计算的所有梯度的总和确定朝向参数的局部最优值的全局下降的大小（这里“本地”是指参数值的空间，并不意味着限于单个参与者）。因此，参与者可以从彼此的训练数据中受益 - 而无需实际看到这些数据！ - 并且可以生成更准确的模型，这些模型本来可以单独学习，仅限于他们自己的训练数据。

参与者可以直接或通过可信的中央服务器交换梯度，甚至可以使用安全的多方计算来“无意地”交换它们，模仿隐藏每个更新来源的可信服务器的功能。出于本讨论的目的，我们假设一个抽象的中心服务器，参与者异步上传梯度。服务器将所有渐变添加到相应参数的值。每个参与者从服务器下载参数的子集并使用它们来更新其本地模型。给定参数的下载标准可以是更新的频率或新近度或添加到该参数的梯度的动态平均值。

**5系统架构**

**5.1概述**

图2说明了我们的协作深度学习系统的主要组件和协议。我们假设有N个参与者，每个参与者都有一个可用的本地私人数据集。所有参与者都提前同意共同的网络架构和共同的学习目标。我们假设存在一个参数服务器，它负责维护所有各方可用的最新参数值。此参数服务器是一种抽象，可以由实际服务器实现或由分布式系统模拟。



每个参与者初始化参数，然后在他自己的数据集上运行训练。该系统包括参数交换协议，该协议使参与者能够将所选择的神经网络参数的梯度上载到参数服务器，并在每个本地SGD时期下载最新的参数值。这允许参与者（i）独立地收敛到一组参数，并且关键地，（ii）避免将这些参数过度拟合到单个参与者的本地训练数据集。一旦网络被训练，每个参与者可以在新数据上独立地和私下地评估它，而不与其他参与者交互。

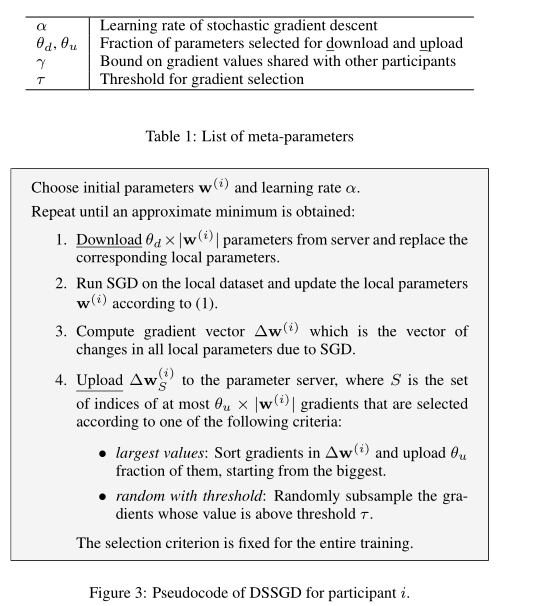
在下文中，我们将描述系统的所有组件。表1列出了我们系统的*元参数*。这些参数控制协作学习过程，而不是正在学习的实际神经网络参数。

**5.2本地培训**

我们假设每个参与者都维持本地神经网络参数的向量。参数服务器维护单独的参数向量,。每个参与者可以随机初始化他的本地参数，或者从参数服务器下载他们的最新值。

然后，每个参与者使用标准SGD算法训练神经网络，在多个时间点上迭代他的本地训练数据。在本地培训期间，不同参与者之间不需要任何协调。它们通过参数服务器间接影响彼此的训练。

图3给出了分布式选择性SGD（DSSGD）算法的伪代码。 DSSGD由每个参与者独立运行，每个学习时期包含五个步骤。首先，参与者从服务器下载部分参数，并用下载的值覆盖他的本地参数。然后，在他的本地数据集上运行一个SGD训练时期。这种训练可以通过一系列小批量进行;小批量是随机选择的大小为M的训练数据点集。



在第三步中，参与者计算，即步骤2中所有参数的变化矢量，即，对于每个参数j，在本地SGD的最后一次运行后从新的值中减去旧的值。我们将称为本地SGD一个时期上的参数j的梯度。值反映每个参数必须改变多少以更精确地模拟第i个参与者的本地数据集。这些信息正是其他参与者需要加入的信息，以避免过度拟合。

有几种方法可以选择在每个本地时代的末端共享哪些梯度。参与者需要就标准达成一致，并在整个DSSGD中始终如一地使用它。我们假设在每个时期最多全部参数里的参数可以选择上传。

我们考虑两个选择标准。第一种方法是选择，选择对梯度下降算法有显着贡献的大值。另一种方法是在大于阈值τ的值中随机选择子集。由于大于τ的梯度的数量可以小于参数的数量，因此将共享更少的梯度。这可能会减慢收敛速度，但这种选择标准更接近我们在扩展具有差异隐私的系统时使用的稀疏矢量技术（参见第7.2节）。

在上传所选择的梯度之前，它们的值被截断为[-γ，γ]范围。为了防止这些值泄漏太多关于训练数据的信息，可以添加随机噪声，如第7节所述。简而言之，参与者用约束（，γ）更新并添加一些在上传之前随机噪声。在第7节中，我们解释了如何设置范围和随机性参数并讨论它们对SGD的影响。

**5.3参数服务器**

参数服务器初始化参数向量，然后处理参与者的上载和下载请求。图4显示了服务器的伪代码。当有人上传梯度时，服务器将上载的添加到相应的全局参数并更新每个参数j的元数据和更新计数器。为了增加最近更新的参数的权重，服务器可以周期性地将计数器乘以衰减因子β，即stat：=β·stat。当参与者从服务器获得具有最大值的参数的最新值时，使用这些统计。每个参与者通过设置θd来决定下载这些参数的哪一部分。

**5.4为什么分布式选择性SGD有效**

我们的分布式SSGD实现了与传统的，侵犯隐私的SGD几乎相同的准确度，原因与SGD成功的原因基本相同：学习过程的*随机性*。在训练期间使用全局参数子集更新本地参数增加了本地SGD的随机性。这在防止当地SGD过度拟合到其较小的本地数据集方面起着至关重要的作用。单独训练时，每个参与者都容易陷入局部最优。使用其他参与者在不同数据集上训练的学习的值来覆盖本地学习参数，帮助每个参与者避免局部最优并使得能够探索其他值，从而产生更准确的模型。

我们的分布式SSGD不会对哪些参数需要被其他参与者更新做出任何假设（？），也不会对更新率做出任何假设。由于更好的计算和吞吐量能力，一些参与者可能经历更多的更新。由于网络错误或其他故障，某些参与者可能无法上传其选定的参数。由于对参数服务器的异步访问，它们还可能会覆盖彼此的更新。竞争条件不仅不会破坏我们的分布式SSGD，事实上它们通过增加随机性来促进其成功。众所周知，由异步参数更新引起的随机性对于训练精确的深度神经网络是有效的[14]。这也与在训练期间随机破坏神经元[47]或输入数据[52]的规则化技术一致，以避免过度拟合。

**5.5参数交换协议**

DSSGD不会假设参与者在上传参数时遵循任何计划。在我们的评估中，我们考虑了以下场景。

对于*循环*，参与者按顺序运行SSGD。每次下载来自服务器的最新参数，运行本地训练，并上传选定的梯度;下一个参与者按固定顺序进行。对于*随机顺序*，参与者以随机顺序下载，学习和上传，但是对服务器的访问是原子的，即参与者在阅读之前锁定它并在写入之后释放锁定。对于*异步*，参与者不协调。虽然一个参与者正在训练一组参数，但其他参与者可以在训练结束之前在服务器上更新它们。

**6评估**

**6.1数据集和学习目标**

我们在深度学习文献中用作基准的两个主要数据集上评估我们的系统。第一个是格式化为32x32图像的手写数字的MNIST数据集[31]，进行归一化，使得数字位于图像的中心。这个数据集由60,000个训练样本和10,000个测试示例组成。

第二个是从Google的街景图像中获得的房屋号码的SVHN数据集[35]。图像为32x32，其中3个浮点数包含每个像素的RGB颜色信息（我们将其转换为YUV）。每个图像都以数字为中心，但许多图像在侧面包含一些干扰物。数据集包含600,000训练图像，我们使用100,000个培训图像，10,000个作为测试样例。表2总结了我们使用的培训和测试示例的数量。

我们通过减去平均值并除以训练集中数据样本的标准差来标准化数据集。 MNIST和SVHN的神经网络输入层的大小分别为1024和3072。学习目标是将输入分类为10个可能的数字之一，因此输出层的大小为10。

**6.2计算框架**

我们使用Torch7 [11,49]和Torch7 nn包。 4这个流行的深度学习库已被Facebook等主要互联网公司使用和扩展。

**6.3神经网络架构**

我们使用两种流行的神经网络架构：多层感知器（MLP）和卷积神经网络（CNN）。 MLP是前馈神经网络架构，其中每层中的神经元完全连接到下一层中的神经元。最初提出反向传播算法来学习这些网络的参数[42]。图1是MLP网络的示例。 CNN是一种特殊的具有稀疏连通性的多层神经网络[31]。 CNN广泛用于图像和视频识别。我们提供了图5（MLP）和图6和7（CNN）中网络架构的确切规格，全部使用Torch7 nn包打印（？）。这些图显示了每层中使用的激活函数（例如，Tanh用于切线双曲线，ReLU用于整流器函数），以及层之间的连接。表3总结了参数的数量。

**6.4实验装置**

我们实现了具有三种不同的参数变换协议的分布式SSGD - 循环，随机顺序和异步。随机顺序的性能与循环非常相似，因此省略。我们将所有结果与两个基线情景进行了比较。第一个是*整个数据集上的Centralized SGD*。这是违反隐私的情况，其中将所有训练数据汇集到一个数据集中，并且使用标准随机梯度下降在该数据集上训练网络。另一种情况是*Standalone SGD。*这是参与者仅在他们的训练数据上训练而没有任何协作的情况。

我们实现了两个标准来选择哪些梯度上传到参数服务器。*使用最大值*，每个参与者上传具有来自上一个本地训练时期的最大绝对值的梯度。*随机随阈值*，参与者上传随机样本的渐变，其值超过阈值。对于*下载*，每个参与者选择已经过最多更新的参数。其他选择标准，例如，下载经历了最大变化的参数，也是可行的。

在所有实验中，参数统计的衰减因子β（参见第5.3节）设定为0.8。我们评估了小批量（1和32）和SGD学习率（α= 0.001和0.01）的几种设置，衰减率为1e-7。我们还将每个DSSGD场景中参与者N的数量在30,90和150之间变化。

我们用整个数据集的1％随机初始化每个参与者的本地训练数据集，即用于MNIST场景的600个数据样本和用于SVHN场景的1000个数据样本。选择用于在SSGD中共享的参数的分数在{1,0.1,0.01,0.001}中取值，即在MNIST上训练MLP的情况下的{15,141,1402,140106}参数（参见表3）。 要下载的参数的分数通常设置为1。

**6.5选择性SGD的结果**

为了显示我们的方法与传统随机梯度下降相比的有效性，我们评估了SSGD和SGD在MNIST和SVHN数据集上训练卷积神经网络（CNN）时的准确性。图8比较了不同元参数值的SGD和SSGD（小批量大小和共享梯度的分数）。一般而言，参与者可以通过训练校准数据集来选择元参数的值，例如，没有隐私含义的公共数据集。

这些结果证实了SSGD背后的直觉：通过在每个梯度下降步骤中仅共享一小部分梯度，我们可以获得与SGD几乎相同的精度。此外，有和没有选择性参数共享的SGD的整体行为是类似的。将小批量大小设置为1在整个训练过程中实现高随机性并且收敛非常快，但也会导致某些曲线的波动。图8显示了直到35次的精度轨迹;除此之外，我们可以实现更高的精度，如表4所示.SSGD可以实现比SGD更高的精度，因为在每个时期仅更新一小部分参数充当正则化技术，它通过防止神经网络权重共同“记住”训练数据来避免过度拟合（这个概念在[47]中描述）。当小批量大小设置为32时，由于在梯度下降期间应用了许多训练数据点的平均梯度，因此收敛更慢但更平滑。

**6.6分布式选择性SGD的结果**

图9显示了在为不同的神经网络架构，参数交换协议和共享参数的分数运行DSSGDon MNIST和SVHN时获得的最佳精度.x轴是共享参数的分数(); y轴是精度，即在测试集中正确分类的数据样本的分数。我们将下载速率设置为1，将学习速率α设置为0.001，将小批量大小设置为32。

在每个图中，我们显示了集中（最大效用，最小隐私）和独立（最小效用，最大隐私）SGD的最佳准确度。两者都独立于x轴，因为两者中都没有参数共享。这两个场景是我们的基线。比较分布式SSGD与基线的准确性反映了效用和隐私之间的权衡。这种差距取决于网络架构，并反映出CNN在训练数据方面与MLP比的优势更大。此外，在我们的设置中，差距受到分类复杂性的影响，每个参与者在SVHN数据集的情况下具有1000个数据样本，在MNIST数据集的情况下具有600个数据样本。

我们的结果表明，任何合作，即使只共享1％的参数，也会产生比独立学习更高的准确性。使用循环参数交换的分布式SSGD具有最高精度，几乎等于集中式SGD。原因是它与SSGD相似（见图8）。为这种准确性付出的代价是学习的速度，这是由最慢的参与者决定的。循环协议适用于所有参与者具有相似计算能力的场景，例如，具有专用SSGD服务器的生物医学研究机构。然而，我们不对本地SGD如何运行做出任何假设。例如，它可以在并行GPU上执行以解决该过程。异步参数交换协议也可以生成准确的模型。其成功的关键在于固有的随机性，因此梯度下降的高随机性可以防止过度拟合。在实施过程中，我们假设参与者可能会落后于其他人，并以概率0.5下载过时的参数集（来自上一个时期的参数）。该协议具有很好的准确性，表明即使在不可靠的（如移动）网络中，DSSGD也能很好地工作。

我们还观察到参与者的数量对准确性的影响小于共享参数的百分比。这表明分布式SSGD不需要很多参与者来提高准确性。

参与者数量，参数更新率和参数交换时间表都会影响分布式SSGD的通信成本。例如，在循环调度中训练具有90个参与者且具有10％的参数上载率的MNIST数据集的MLP模型需要服务器每次需要上载90×14010×32 = 38.5兆字节的参数参数下载速率为100％，服务器需要在每个时期支持385兆字节的下载。

假设每个参与者与其他参与者共享最大的梯度，则获得所有上述结果。另一种方法是从其值超过阈值的梯度中随机采样。图10比较了MNIST和SVHN数据集的DSSGD与这两个标准的准确性。在“随机阈值”场景中，我们首先将梯度值Δw截断为[-0.001,0.001]范围，然后按随机顺序进行处理，并在abs（Δwj）≥τ时上传。在所有实验中，神经网络结构（CNN），学习率（α= 0.001），小批量（M = 1）和交换协议（循环）都是相同的。在“随机阈值”场景中，可以上载梯度少于θu分数，因此准确度有时较低。为了找到阈值τ的有效值，参与者需要在公共校准数据集上运行DSSGD。

图11显示了DSSGD对不同数据集，学习率和参与者数量的收敛性。上传速率θu为0.1，下载速率θd为1，小批量大小为32，参数交换协议为循环，梯度选择标准为最大值。这些结果表明，无论参与者的数量如何，较高的学习率都会导致更快的收敛到最大的准确度。因此，DSSGD的分布式和选择性不会改变梯度下降算法的整体行为。

**7隐私**

我们的系统旨在解决与深度学习相关的若干隐私威胁。首先，在传统的深度学习中，所有培训数据都透露给第三方（通常是执行学习的公司），而贡献数据的个人对其没有任何控制权。他们的敏感信息可能会泄露给公司本身，泄露给破坏公司数据存储的攻击者，以及可以通过法律和法律手段访问数据的执法和情报机构。

其次，在传统的深度学习中，数据所有者无法控制学习目标（即，正在训练哪个模型），因此无法控制甚至不知道从他们的数据推断出什么。例如，个人可能愿意将她的图像分享用于面部识别，而不是用于从背景推断她的位置。

第三，在传统的深度学习中，学习的模型不能直接为数据所有者所用。如果他们想要使用它，他们必须向持有该模型的公司显示他们的输入，从而使他们暴露于与训练数据相同的隐私风险。

我们的隐私保护深度学习系统解决了所有这些问题，旨在保护培训数据的隐私，确保公众了解学习目标，保护应用学习模型的数据的隐私，以及模型输出的隐私。

我们考虑的场景 - 例如，医学机构之间的图像识别模型的协作学习 - 涉及非主动恶意的参与者。因此，假设“被动”对手模型是合理的，其中参与者按照设计执行协议，但可以尝试学习或者从其他参与者的数据中提取敏感信息。

**7.1防止直接泄漏**

*在训练模型的同时*。与传统的深度学习不同，在我们的系统中，参与者不会向任何人透露他们的训练数据集，从而确保他们数据的强大隐私。本地数据集的大小和动态是保密的，每轮SSGD可以使用不同的数据样本。参与者还可以随时删除他们的训练数据。

*在使用模型时*。所有参与者都可以学习该模型，并且可以在本地和私下使用它，无需与其他参与者进行任何通信，也无需向任何人透露输入数据或模型的输出。因此，与传统的深度学习相比，使用该模型时绝对没有泄漏。

**7.2防止间接泄漏**

我们系统的参与者可以通过在训练期间对神经网络参数的一小部分进行公开更新，间接地揭示他们的训练数据集的一些信息。每个参与者完全控制分享哪些梯度，并可能决定不分享特别敏感的梯度。此外，每个参与者只分享他的梯度的一小部分：正如我们所示，即使共享少数1％仍然导致比学习本地数据明显更好的准确性。即便如此，我们使用差异隐私来确保参数更新不会泄漏有关训练数据集中任何单个点的过多信息。

***差异隐私***。我们对差异隐私参数更新的应用受到最近关于隐私保护经验风险最小化的研究的启发[4]。简而言之，如果产生给定输出的概率不太依赖于特定数据点是否包含在输入数据集中，则计算是差异私有的[19]。对于任何两个数据集D和D'在单个项目中的不同以及函数f的任何输出O.，Pr {f（D）∈O}≤exp（e）·Pr {f（D 0）∈O}。 （2）(?)

参数e控制不同私有f的准确性与泄漏的信息量之间的权衡。（不同数据集，参数尽量相同）

在我们的例子中，f计算参数梯度并选择与其他参与者共享的参数。有两个潜在的潜在泄漏源：如何选择梯度进行共享以及共享梯度的实际值。为了减轻这两种类型的泄漏，我们使用**稀疏矢量技术**[20,25]来（i）随机选择其值大于阈值的梯度子集，并且（ii）共享所选梯度的扰动值，所有这些都在一个一致的差异私人机制下。这相当于将响应发布到其值大于公知阈值的查询。

让每个参与者的每个DSSGD时代的总隐私预算为e.我们将这个预算分成c部分，其中c是我们可以在每个时期上传的总梯度数（即，c =θu|Δw|）。然后将每个潜在上传的预算分为两部分。第一个将用于检查随机选择的参数j的梯度是否高于阈值τ。如果梯度高于阈值，则第二个将用于实际释放（上载）梯度。我们使用拉普拉斯机制在选择期间添加噪音，并根据分配的隐私预算上传。噪声取决于隐私预算以及每个参数的梯度敏感度。在下文中，我们假设所有参数具有相同的灵敏度Δf，但这不是必需的，并且不同的参数可能具有不同的敏感度。

图12显示了差分私有DSSGD的伪代码。为了分割e，我们遵循[20] .的致力于选择，其中部分用于随机噪声，另一部分用于随机噪声r τ。剩余的用于释放的值。注意在失败的阈值检查后不再重新生成rτ。这不仅确保所有共享梯度都是差异私有的，而且还保证不会为与其他参与者共享的小梯度支付隐私“惩罚”。

***估计灵敏度***。函数的灵敏度决定了需要将多少随机噪声添加到其输出以实现相应的隐私。 f的（全局）灵敏度是

Δf= max D，D 0 || f（D）-f（D’）||。 （3）

估计随机梯度下降的真实敏感性具有挑战性。相反，我们修改函数，使其输出保持在固定的，与输入无关的边界，并使用这些边界来估计灵敏度：这是对可能与其他参与者共享的梯度值强制执行[-γ，γ]范围的约束函数（第5节）。这种方法可能会降低准确性（虽然在我们的案例中效果可以忽略不计），但保证了隐私。以前使用类似技术来强制使用不受信任的映射器来实现MapReduce计算的隐私[40]。

通过帮助避免过度拟合，限制参数和梯度的值范围甚至可以改善训练过程。一些现有的正则化技术已经迫使对参数的范数进行约束。 Max-norm已被用于协同过滤[46]和深度学习[47]。此外，截断到[-γ，γ]范围内的梯度值表示梯度下降期间移动的方向和大小。因此，小的γ值（意味着更小的灵敏度，因此更小的噪声和更高的准确度）将影响算法的学习速率，而不是最佳解决方案是否可实现。此外，当聚合多个参与者的梯度时，梯度下降算法可以遍历局部最优。我们讨论扰动对分布式选择性SGD的影响。

元参数γ独立于训练数据而设置，因此不会泄漏任何敏感信息。可以通过对校准数据集进行训练来设置，其输入与实际输入类似，但不是隐私敏感的。然后我们（过度）估计我们算法的灵敏度为2γ并将上传的梯度截断为[-γ，γ]范围。这有助于减轻非常大的噪声值对训练过程的不利影响。

我们预计，通过确保所有梯度的范数在全局或局部（例如，穿过通向给定神经网络节点的所有边缘）每次更新时都有界，可以显著降低全局灵敏度估计，从而提高精度。事实上，后一种规范边界是一种已知的正规化技术。我们计划在未来的工作中研究规范边界到差异私人深度学习的应用。

**7.3实验结果**

我们评估不同值的影响？ （差分隐私参数），N（参与者的数量）和θu（上传的梯度的分数）对使用具有差分隐私的分布式选择性SGD的神经网络的精确度。

图13显示了结果，并将它们与独立学习和集中式SGD进行了比较。我们将约束γ设置为0.001，将阈值τ设置为0.0001。正如所料，更小？值（即，更强的差异保密）导致较低的准确性。然而，对于许多参与者以及当参与者共享其大部分梯度时，差异私有DSSGD的准确性优于独立训练的准确性。

**7.4不经意的参数服务器**

无论参数服务器是否可信，培训数据分离和差异隐私的隐私保护仍然存在。但是，为了防止好奇的服务器链接每个参与者的更新，可以设计一个不考虑上传者身份的参数服务器。例如，参与者可以匿名验证自己和他们上传的渐变[7]。具有可证明安全性的可扩展匿名通信协议可用于隐藏参与者的身份[12,54]。

分布式SSGD中参数彼此独立，这是基础随机梯度下降算法所固有的，也使参数存储系统的完全分布式实现成为可能，其中每个参与者对参数的随机子集负责。我们将该方案的详细设计留给未来的工作。

**8结论**

这项工作是将隐私带入机器学习方法的第一步，这种方法正在彻底改变人工智能。我们提出了一种基于选择性随机梯度下降的新型分布式训练技术。我们的方法适用于任何类型的神经网络，并保留参与者训练数据的隐私，而不会牺牲所得模型的准确性。因此，它可以帮助将深度学习的好处带到数据所有者无法通过机密性问题共享数据的领域。

致谢。我们感谢Adam Smith解释如何在我们的环境中应用稀疏矢量技术和其他差分隐私机制。 NSF拨款1223396和1408944，美国国家医学图书馆的NIH拨款R01LM011028-01和瑞士国家科学基金会的博士后研究员RezaShokri部分支持了这项工作。