强化学习：提高交流效率的策略

雅的的的。布伦丹·麦克马汉，菲利克斯X. Yu，阿南达，苏雷什和戴夫·培根č*
谷歌
{康基，麦克马汉，费利克西尤，他们，达巴}@google.com

皮特里克á
阿卜杜拉国王科技大学（KAUST），图瓦尔，沙特阿拉伯大学，爱丁堡大学，爱丁堡，苏格兰
皮特Richtarik@kaust .edu .sa

一个分心

联邦学习是一种机器学习设置，其目标是训练一个高质量的集中式模型，同时训练分布在大量客户端上的数据维护，每个客户端都有不可靠和相对缓慢的网络连接。在每一轮中，每一个客户依赖于其本地数据计算当前模型的更新，并将此更新通信到中央服务器，在中央服务器上更新重新聚合以计算新的全局模型。在这种情况下，典型的客户是手机，而通信效率则至关重要。
在本文中，我们提出了两种降低上行通信成本的方法：
结构化更新，我们直接从一个限制的spa参数化少量的变量，e.g.either低秩或随机掩码；和草图更新，我们学习一个完整的模型更新，然后使用量化、随机旋转和连接到服务器的子采样的组合按它。卷积网络和递归网络的实验结果表明，该方法可以将通信成本降低两个数量级。

1i定向

随着数据的增长和模型的复杂，训练机器学习越来越需要将模型参数的优化分布在多个平台上。存在-线学习算法是为高度控制的环境（如数据中心）设计的，其中数据分布在balancedandi.i.d中的机器之间。时尚和高速电脑网络是可用的。
最近，联邦学习（以及相关的分散方法）(McMahan&Ramage，2017;
Kone n等人，‘y2016;麦克马汉等人，2017;Shokri& Shmatikov ,已被提议为
另一种选择：一个共享的全局模型在一个中央服务器的协调下，从参与设备的联邦政府中进行训练。参与设备（客户端）通常数量多，互联网连接低或不稳定。当训练数据来自用户与移动应用程序的交互时，虚拟人学习的一个主要激励例子就出现了。联邦学习使手机能够在协作学习一个共享的预测模型的同时，将所有的训练数据保存在设备上，将进行机器学习的能力与存储数据的需要解耦。训练数据通常隶属于用户的移动设备，这些设备被用作节点，对其本地数据进行计算，以更新全局模型。这不仅仅是使用本地模型来对移动设备进行预测，还可以为移动设备带来模型训练。上述框架不同于传统的分布式机器学习(Reddi等人，2016;马等人，2017;沙米尔等人，2014;张林2015;院长

*同时隶属于爱丁堡大学时完成的工作。

以及其他人2012;奇林比等人，由于客户数量众多，高度不平衡和non-i.i.d.在每个客户端上都有可用的数据，以及相对较差的网络连接。在这项工作中，我们关注的是最后一个约束，因为这些不容易的和不对称的联系对实际的联邦学习提出了一个特殊的挑战。

为了简单起见，我们考虑了联邦学习的同步算法，其中一个典型的回旋算法由以下步骤组成：

1.选择现有客户端的一个子集，每个客户端下载当前模型。
2.子集中的每个客户端都根据其本地数据计算一个更新后的模型。
3.模型更新将从选定的客户端发送到服务器。
4.服务器聚合这些模型（通常是通过平均），以构建一个改进的全局模型。

上述框架的简单实现要求每个客户在每一轮中将一个完整模型（或完整模型更新）发送回服务器。对于大型模型，由于多种因素，这一步骤很可能是联合学习的瓶颈。一个因素是互联网连接速度的不对称特性：上行通常比下行慢得多。美国的平均宽带速度是下载为55.0Mbps，上传为18.9Mbps，一些互联网服务提供商明显更加不对称，例如，Xfi速度为125 Mbps。15 Mbps以上(速度测试。网2016).此外，现有的模型压缩模式，如ha1。（2015）可以减少频带-下载当前模型所需的宽度，以及加密协议，以确保单个客户端的更新可以在平均使用数百或数千个其他更新之前进行检查(Bonawitz等人。进一步增加需要上传的比特数的数量.

因此，研究能够降低上行通信成本的方法是很重要的。在本文中，我们研究了两种一般的方法：

结构化更新，我们直接从限制的空间学习更新，可以参数化更少的变量。
草图更新，在那里我们学习一个完整的模型更新，然后在发送到服务器之前共同按它。

这些方法，将在第2节中详细解释和3，可以组合起来，例如，首先学习a结构化更新和草图；我们在这项工作中没有尝试这种组合。在下面，我们将正式地描述这个问题。联邦学习的目标是学习一个参数体现在实矩阵1中的模型 **W∈Rd1×d2从数据存储在大量的客户端。我们首先提供了联邦学习的一种幼稚的交流版本。在sount≥0中，服务器将当前模型Wt分配到nt客户端的子集St。这些客户端会根据它们的本地数据独立地更新模型。让更新后的本地模型进入**
是W 1 t，W 2 t，…所以客户端i的更新可以写成∈：=∈-∈t，对于∈。
这些更新可以是在客户机上计算的单个梯度，但通常会是更复杂的计算的结果，例如，在客户机的本地数据集上进行的随机梯度下降（SGD）的多个步骤。在任何情况下，每个选定的客户端然后将更新发送回服务器，在其中全局更新是通过聚合2来计算的所有客户端更新：

$$\mathbf{Wt+1= W\ t+\eta\ t\ H\ t,\ Ht := \varepsilon\ i\in\ St\ H\frac{1}{n_t}i}.$$

服务器选择学习rateη t。为简单起见，我们chooseη t= 1。

在第4节中，我们描述了神经网络的联邦学习，其中我们使用了一个单独的二维空间矩阵W表示各层的参数。我们假设W得到右乘，即d1和d2分别表示输出和输入维数。注意，一个完全连接层的参数自然地表示为2个d矩阵。然而，卷积层的内核是一个形状#输入×宽度×高度×#输出的四维张量。在这种情况下，W从内核重塑到形状（#输入×宽度×高度）×#输出。

概述和总结。提高联邦学习的沟通效率的目标是
<u>为了降低将H i t发送到服务器的成本，同时从存储在大量数据中的数据中学习</u>

为了简单起见，我们只讨论单个矩阵的情况，因为一切都延续到设置
具有多个矩阵，例如对应于一个深度神经网络中的单个层。2一个加权和可以用来代替基于
特定i模型的平均值。

互联网连接和计算可用性有限的设备。我们提出了两类一般的方法，结构化更新和草图更新。在实验部分，我们评估了这些方法在训练深度神经网络中的效果。

在对CIFAR数据进行的模拟实验中，我们研究了这些技术对联邦平均算法的收敛性的影响(McMahan等人，2017).由于收敛速度只有轻微的退化，我们就能够将通信的数据总量减少两个数量级。这使我们可以使用全卷积模型获得良好的预测精度，而总通信的信息比原始cifar数据的大小要少。在一个对用户划分的文本数据进行的更大的现实实验中，我们证明了我们能够在使用每个用户的数据一次之前，有效地训练一个递归神经网络来进行下一个单词预测。最后，我们得到了最好的结果，包括对具有结构化随机旋转的更新进行预处理。这一步的实际效用对于我们的设置是独特的，因为在典型的SGD并行实现中，应用随机操作的成本将占主导地位，但与联邦学习中的局部训练相比，这是不可行的。

2个结构化的U PDA TE

第一种类型的通信高效更新限制更新必须预先指定
构造本文考虑了低秩掩模和随机掩模两种结构。需要强调的是，我们直接训练这种结构的更新，而不是用特定结构的对象进行近似配对/绘制一般更新——这将在第3节中讨论。

大多数k，其中k是固定数。按顺序　　要这样　　我们表示为两个矩阵的乘积：
低等级。我们强制对局部模型Hit∈Rd1×d2的每一次更新都设为秩在的低秩矩阵
它，其中∈1×k，=∈Rk×d2 在随后的计算中，我们生成了A i t
在局部训练过程中随机考虑一个常数，我们只优化位。笔记

在实际实现中，它可以以随机种子的形式被压缩
客户端只需要发送训练的B到服务器。这种方法立即节省了一个因素
d 1/k不通信。我们在每一轮和为每个客户端重新生成矩阵A i t
独立地

我们还尝试固定B和训练A，同时训练A和B；没有表现
好我们的方法似乎与Denil等人所考虑的最佳技术一样好。(2013),不需要任何手工制作的功能。对这一观察结果的一个直观的解释是
接下来的我们可以解释b是投影矩阵，而a是重构矩阵。固定A
而对B-i的优化则类似于问“给定一个给定的随机重建，什么是投影”
这将恢复最主要的信息吗？”。在这种情况下，如果重构是全秩的，则存在恢复由顶部关键向量所跨越的空间的投影。然而，如果我们随机固定投影并寻找一个重建，我们可能是不幸的，重要的子空间可能已经被投影出来，这意味着没有重建将做得尽可能好，或者将非常难以得到。

随机掩模。我们限制更新命中为一个稀疏垫矩阵，遵循预先定义的随机
稀疏性模式（即，一个随机掩模）。该模式在每一轮中重新生成，并在每个客户端中独立生成。与低秩方法类似，稀疏模式可以完全由
一个随机的种子，因此它只需要发送命中的那个-零入口的值，
还有种子。

3个的了

第二种类型的更新解决通信成本，we调用素描，首先计算
在没有任何约束的局部训练中的完全命中，然后近似或编码
在发送到服务器之前，以（有损）压缩形式进行更新。服务器在进行聚合操作之前解码更新内容。这种草图方法有在许多领域的应用(伍德拉夫，2014).我们实验使用多个工具来执行草图，这些工具是相互兼容的，可以联合使用：

下采样除了发送命中之外，每个客户端只通信已形成的矩阵^Hit
从H到它的（缩放）值的随机子集。然后，服务器对下采样的更新进行平均，
生成全局更新^ H t。这样做可以使样本d更新的平均值是真实平均值的无偏估计量： E[^ H t] = H t。与随机掩码结构更新类似，掩码在每一轮的每个客户端都是独立随机的，掩码本身可以作为同步种子存储。

概率量化。另一种压缩更新的方法是通过量化权重。

我们首先描述了将每个标量量化到一比特的算法。考虑一下更新的点击量，让我们来吧
h=（h 1，...，hd1×d2）=~vec（Hit），让最大=最大j（h j），最小=最小j（h j）。压缩的
h的更新，用h表示，生成如下：

$$\sim h_j = \begin{cases} h_{max}, & \text{带概率 } \frac{h_j - h_{min}}{h_{最大} - h_{最小值}} \\ h_{min}, & \text{带概率 } \frac{h_{最小值}}{h_{最大} - h_{最小值}} \end{cases}.$$

证明~h是h的无偏估计。该方法提供了32×的压缩量
与4个字节数相比。分析了该压缩方案所产生的误差
Suresh等人的实例。（2017），是在中提出的协议的一个特殊情况我和我的朋友
(2016).

我们还可以将上述结果推广为每个标量的超过1位。对于b位量化，我们首先将[hmin，hmax]等分为2b区间。假设hi落在以hI和h II为界的t个区间内。量化的操作是分别用hI和h II替换上述方程的hmin和h II。参数b允许以简单的方式平衡精度和通信成本。

另一种量化方法也来自于在平均变量时减少通信。 (2016).增量、随机和分布的时间化算法可以在一个量化的更新设置中类似地进行分析(Rabbat& No wak，2005;戈洛文等人，2013;加马尔和莱，2016).

通过结构化的随机旋转来改进量化。当不同维度的尺度近似相等时，上述1位和多位泉化方法效果最好。例如，当最大=1和最小= -1且大多数值为0时，1位量化将导致很大的误差。我们注意到，在量化之前对h应用一个随机旋转（乘以h bya随机正交矩阵）解决了这个问题。这一主张在苏雷什等人身上得到了理论上的支持。(20 17).在该研究中，结构随机旋转可以将量化误差减少O（d/logd），其中d为h的维数。我们将在下一节中展示它的实际效用。

在解码阶段，服务器需要在聚合所有更新之前执行逆旋转。请注意，在实践中，ofh的维数可以很容易地高达d= 106或更多，这在计算上无法生成（O（d 3））并应用(O（d 2）一般的旋转矩阵。和苏雷什等人一样。（2017），我们使用一种结构化旋转矩阵，它是a的乘积沃尔什-阿达玛矩阵和二元对角矩阵。这降低了生成和将矩阵应用于O(d)和O（dlog d）的共同计算复杂性，这与联邦学习中的局部训练相比是微不足道的。

4个实验

我们使用联邦学习来训练深度神经网络来完成两种不同的任务。首先，我们实验了CIFAR-10图像分类任务(克里热夫斯基， 2009)与卷积网络和人工划分的数据集，并详细探讨了我们提出的算法的性质。其次，我们使用更现实的场景来进行联邦学习——公共编辑编辑帖子数据（谷歌BigQuery），来训练一个循环网络的下一个单词预测.

Reddit数据集对于模拟F扩展学习实验特别有用，因为它包含自然的每个用户的数据分区（由文章的作者编写）。这包括了在实际实施中预期会出现的许多特征。例如，许多用户拥有相对较少的数据

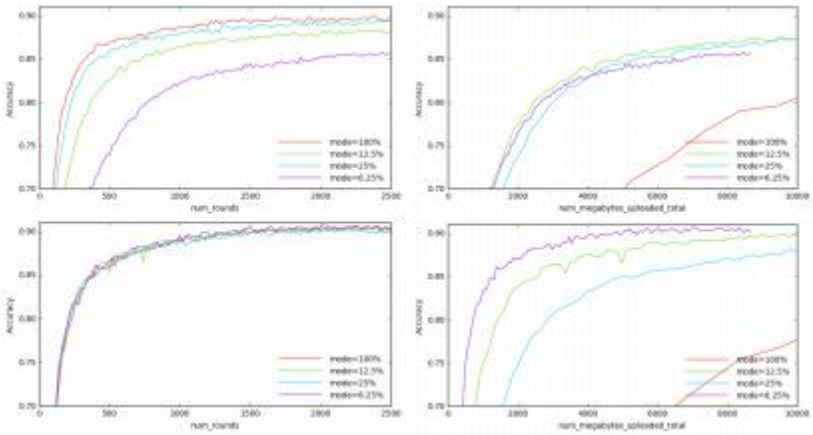

图1：使用CIFAR数据的结构化更新，以缩小各种模式。低秩更新在顶部行，随机掩码更新在底部行。

大多数用户使用的点和单词都围绕着特定用户最喜欢的特定主题聚集。

在我们所有的实验中，我们都使用了联邦平均算法(McMahan等人， 2017),这显著减少了训练一个好的模型所需的c通信的轮数。然而，我们希望我们的技术在应用于同步分布式SGD时，能够显示出类似的通信成本降低，见实例Alistarh等人。(2016).联合平均，在每一轮中均匀随机选择多个客户端，每次在其本地数据集上执行几个SGD周期，学习速率为η。为结构化
更新时，SGD只被限制在受限制的空间内进行更新，即只有位的条目
对于低秩更新和未掩码条目的随机掩码技术。从这个更新
模型，我们计算每一层命中的更新。在所有情况下，我们用一个范围进行实验
选择的学习率，并报告最好的结果。

4.1c-10d的版本

在本节中，我们使用cifar-10数据集来检验我们所提出的方法的属性，作为联邦平均算法的一部分。

在CIFAR-10数据集中有50000个训练示例，我们将其随机划分为
100个客户，每个客户包含500个培训例子。我们使用的模型架构是从sprineberg等人描述的全卷积模型。（2014），适用于总共超过106个参数。虽然这个模型不是最先进的，但它足以满足我们的需要，因为我们的目标是评估我们的压缩方法，而不是在这个任务上达到最好的准确性。

该模型有9个卷积层，第一个层和最后一个层的参数rs比其他的少得多。因此，在这整个部分中，当我们试图减少单个更新的大小时，我们只压缩内部的7层，每个层具有相同的参数3。对于所有的方法，我们用关键字“模式”来表示这个问题。对于低秩的更新，“模式=25%”指的是更新的秩设置为全层变换的秩的1/4，对于r andom掩模或草图，这指的是除25%之外的所有参数被归零。

在图1中总结的第一个实验中，我们比较了两种类型的结构化更新第2节介绍-低排名的顶部一行和随机掩码在底部一行。主要信息是，随机掩码的性能明显优于低秩，因为我们减少了更新的大小。特别是，随机掩模的收敛速度似乎在本质上不受影响

3我们也尝试减少所有9层的大小，但这在通信中产生的感知可以忽略不计，同时也略微降低了收敛速度。

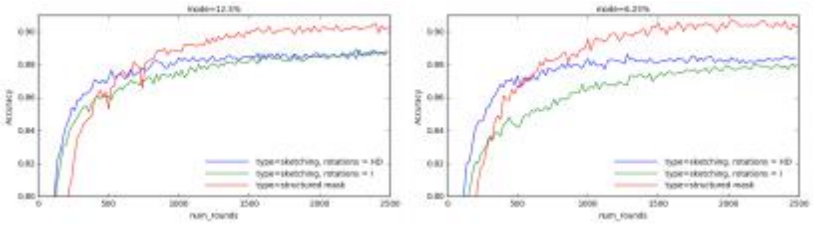

图2：对CIFAR数据进行结构化随机面具更新和统计未量化的日期的比较。

如果以轮数的万亿来衡量。因此，如果目标仅仅是只最小化上传大小，那么减少更新大小的版本将是明显的赢家，如右栏所示。

在图2中，我们比较了结构化更新和草图更新的性能，但没有任何量化-tion.由于在上面的内容中，结构化随机掩更新表现得更好，为了清晰，我们省略了低秩更新。我们将其与这些更新数据的性能进行比较，使用并不使用随机旋转对更新进行预处理，如第3节所述，以及两种不同的模式。我们用“HD”表示随机的阿达玛旋转，用“I”表示名词符号。

直观的期望是，直接学习结构化随机掩码更新应该比学习非结构化更新更好，非结构化更新是用相同参数数量的n个草表示。这是因为通过画草图，我们抛弃了一些在训练过程中所包含的信息。事实上，通过绘制更新，我们应该接近到一个略低的精度，理论上可以支持，使用类似的论证，如在(Al-伊斯塔等人，2016年)，因为进行更新增加了直接出现在转换中的方差-收敛性分析。当使用结构化随机掩码更新时，我们可以看到这种行为，我们最终能够收敛到略高的精度。然而，我们也可以看到，通过绘制更新的草图，我们能够以稍快的速度获得适度的精度（例如，85%）。

在最后一个关于CIFAR数据的实验中，我们重点关注了在第3节中引入的所有三个元素的相互作用-子采样、量化和随机旋转。请注意，所有这些工具的组合将实现比上述实验更高的压缩率。图3中的每一对绘图专注于特定的模式（子采样），并在每次在量化中使用不同的位，有或没有随机旋转。我们可以在所有的图中一致地看到的是，随机旋转提高了性能。一般来说，该算法的行为在没有旋转的情况下不太稳定，特别是在少量的量化比特和较小的模式的情况下。

为了突出公共阳离子储蓄的潜力，注意通过预处理随机旋转，绘制除了6.25%的更新和使用2位量化，我们只得到一个小的下降收敛，而节省256的位需要代表个人层的更新。最后，如果我们对最小化上传的数据量感兴趣，我们可以获得一个适度的准确性，比如85%，而总的通信量低于
上传原始数据所需费用的一半。

4.2对它的预测

我们基于Reddit（谷歌BigQ uery）上公开发布的帖子/评论的数据，构建了模拟联邦学习的数据集，如前所述Al-Rfou等人。(2016).对于我们的目的来说，关键的是，数据库中的每一篇文章都由作者键控，所以我们可以通过这些键对数据进行分组，从而假设有一个客户端设备的伪作者。有些作者有大量的帖子，但在每一轮的FedAvg中，我们最多处理32 000个代币超级用户。我们省略了具有少于1600个令牌的作者，因为在模拟中每个客户机的开销是恒定的，而数据很少的用户对训练没有太大贡献。这样剩下763430个用户的数据集，平均每个用户有24791个令牌。为了进行评估，我们使用了一个相对较小的测试集，由75 122 个标记组成。

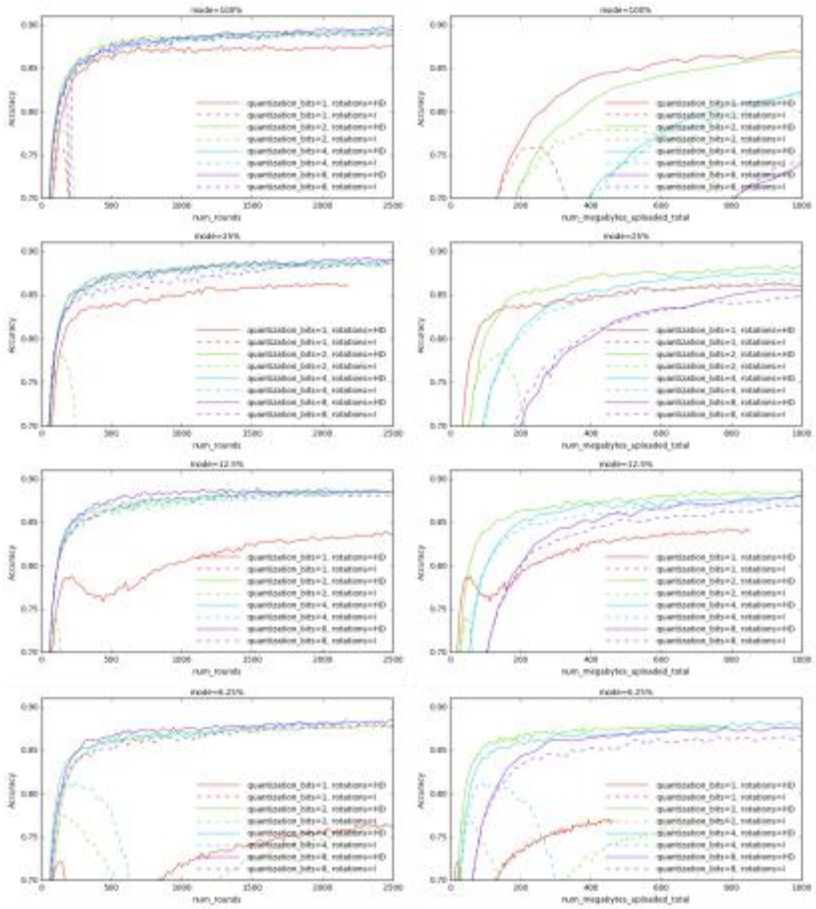

图3：草图更新的比较，将预更新准备与CIFAR数据的旋转、量化和子采样相结合。

基于此数据，我们训练了一个LSTM下连词预测模型。该模型被训练来预测下一个单词，给定当前的单词和从前一个时间步长传递的状态向量。该模型的工作原理如下：通过在一个包含10 017个单词（标记）的字典中查找单词，将单词st映射到一个嵌入向量et∈R 96。然后，et与模型在前一个时间步长s t1∈R 256中发出的状态组成，发出一个新的状态向量st和一个“输出嵌入”

o t∈ R 96 .输出嵌入通过内积对词汇表中每个项的嵌入进行评分，然后通过softmax进行归一化，以计算词汇表中的概率分布。与其他标准语言模型一样，我们将每个输入seqnce视为以隐式的“BOS”（序列的开始）标记开始，以隐式的“EOS”（序列的结束）标记结束。与标准的LS TM语言模型不同，我们的模型对嵌入层和软极大层使用相同的学习嵌入。这将模型的尺寸减少了约40%，以略微降低模型质量，这是移动应用程序的一个优势。许多标准LSTMRNN方法的另一个变化是，我们训练这些模型限制单词嵌入，使其具有固定的l2范数为1.0，这可以提高收敛时间。该模型总共有1.35 M的参数。

为了减少更新的大小，我们绘制了所有的模型变量，除了一些小的变量（如偏差），它们消耗的内存小于0.01%的内存。我们使用AccuracyTop1来评估模型分配最高概率的单词正确的概率。我们总是把它算为一个错误，如果真正的下一个词是没有的字典，即使模型预测“未知”。

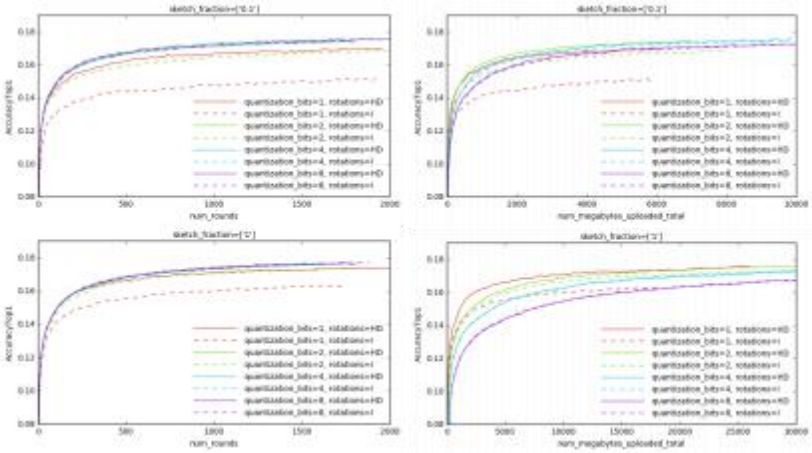

图4：草图更新的比较，在Reddit数据上训练一个循环模型，每轮随机抽样50个客户端。

在图4中，我们在Reddit数据上运行联邦平均算法，使用各种参数 指定Setching。在每次迭代中，我们随机抽取50个用户，根据本地可用的数据计算更新，绘制草图，并取所有更新的平均值。抽样10；20，每轮100个客户提供了类似的结论。

在所有的图中，我们结合了这三个组成部分来绘制在步骤3中引入的更新。首先，我们应用随机旋转来预处理局部更新。此外，“粗略分数”设置为0.1或1，表示被下采样的更新元素的分数。 _

在左栏中，我们将其与迭代的次数相对起来。首先，我们可以看到，随机旋转的预处理效果具有显著的正效应，特别是对于小数量的量化位。有趣的是，对于所有的选择采样比，随机的阿达玛转换量化为2位不会造成任何性能损失。突出显示的一个重要指标是在图中显示的轮数是
2000 .因为我们每轮抽取50个用户，所以这个实验甚至一次都不会接触到大多数用户的数据！这进一步加强了在现实环境中应用联邦学习是可能的，而不会以任何方式影响用户体验的主张。

在右列中，我们将相同的数据与客户端需要通信回服务器的总兆字节数绘制出相同的数据。从这些图来看，如果我们需要彻底地最小化这个度量，我们提出的技术是非常有效的。当然，这些目标都不是我们在实际应用中优化的。然而，鉴于在联邦学习的大规模部署中固有的问题上目前缺乏经验，我们认为这些都是与实际应用相关的有用代理。

最后，在图5中，我们研究了在单轮中使用的客户数量对服务器的影响 - 应力双折射我们对固定的轮数（500和2500轮）运行联邦平均算法，每轮的客户端数量不同，将更新量化到1位，并绘制得到的精度。我们看到，每轮有足够的客户数量，在这种情况下为1024，我们可以将下采样元素的比例降低到1%，与10%相比，准确率只有轻微的下降。这是联邦设置中一个重要和实用的权衡：人们可以在每个周期中选择更多的客户端，同时让它们通信更少（例如，更积极的子采样），并使用更少的客户端获得相同的准确性，但每个客户端通信更多。当许多客户端可用时，格式可能是更好的，但每个客户端的上传带宽都非常有限——这是实践中的一种设置。

R EFERENCES

拉福，马克皮克特、哈维尔斯奈德、宋云珊、布莱恩斯特罗普和雷库兹韦尔。Co nversational
上下文线索：个性化和历史上的反应排名的案例。arXiv:1606.0037 2, 2016.

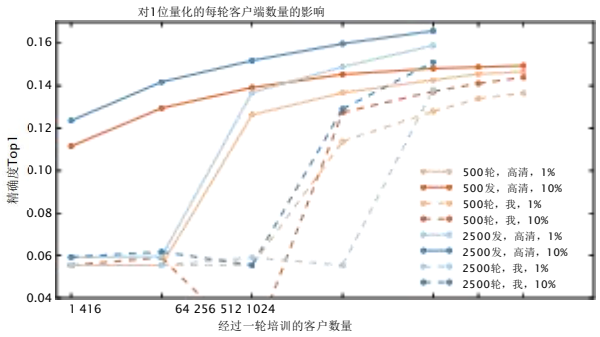

图5：每轮培训中使用的客户数量的影响。

丹阿里斯塔，李，富冈良，米兰沃伊诺维奇。QSGD：通信最优随机梯度下降的随机量化。arXiv:1610.02132, 2016.

基思·博纳维茨、弗拉迪米·伊万诺夫、本·克鲁特、安东尼奥·马塞多内、布伦丹·麦克马汉、萨瓦尔·帕特尔、丹尼尔·拉梅奇、亚伦·西格尔和卡恩塞斯。隐私保护机器学习的实用聚合。InACM计算机与通信安全会议（ACM中国化学会），2017年。

奇林比，苏祖，约翰逊，卡蒂克和卡利亚纳拉曼。项目开发公司：构建高效、可扩展的深度学习训练系统。在第11届USENIX操作系统设计与实现研讨会（OSDI 14）上，页。571–582, 201 4.

杰弗里·迪恩，格雷格·科拉多，拉贾特·蒙加，陈凯，马修·德文，马克，老安德鲁，保罗·塔克，柯杨，引用V Le等。大规模的分布式深度网络。在NIPS中，第1223-1231页，2012年。

米沙丹尼尔，巴巴克沙基比，洛朗丁，南多德弗雷塔斯，等。在深度学习中的预测参数。在NIPS中，第3页。2148–2156, 2013.

穆斯塔法和李丰黎。在随机分布和坐标下降的量化更新。arXiv:1609.05539, 2016.

丹尼尔·戈洛夫，D.斯卡利，H.布伦丹·麦克马汉，和迈克尔·杨。大规模学习与较少的分支随机化。在ICML，2013年。

谷歌大查询。Reddit评论数据集。大查询，2016年。https://bigquery.clo ud。谷歌。com/dataset/fh-bigquery .

宋汉、毛惠子、威廉·杰帝力。深度压缩：用剪枝、训练量化和霍夫曼编码压缩深度神经网络。arXiv预印本，arXiv：1510.00149,2015年。

雅库布·肯恩伊和彼得·里克特里克。随机分布均值估计：准确性与通信。ča *arXiv:1611.07555, 2016.*

雅库布恩，布伦丹麦克马汉，丹尼尔拉梅奇和彼得里奇t里克。联邦优化：对设备上智能的机器学习。arXiv预印本，arXiv：1610.02527,2016年。 č á

亚历克斯·克里耶夫斯基。从微小的图像中学习多层特征。技术报告，2009年。

陈欣，国王，贾吉，弗吉尼亚史密斯，迈克尔一世乔丹，彼得里克特里克，和马丁čá德。具有任意局部求解器的分布式优化。优化方法与软件，32(4)：813-848,2017。áč

H.，布伦丹·麦克马汉南和丹尼尔·拉梅奇。联邦学习：使用超集中训练数据的协同机器学习。的研究。googleblog .联邦学习协作。html, 2017.

H.，布伦丹·麦克马汉，艾德尔·摩尔，丹尼尔·拉马奇，赛斯·汉普森，布莱斯·阿吉拉和阿卡斯。从分散的数据中获得深度网络的通信高效学习。发表在2017年第20届人工智能与统计学（AISTATS）国际会议论文集上。

M.G.拉巴和R.D.诺瓦克。用于分布式数据优化的量化增量算法。IEEE关于通信选定领域杂志，23(4)：798-808,2005。

雷迪，凯恩，彼得里奇特里克，巴纳布斯，和亚历克斯斯莫拉。快速、通信、有效的分布式优化。arXiv:1608.06879, 2016.čáá66

沙米尔、内森·斯雷布罗和张童。使用近似牛顿型方法的通信高效分布式优化。在ICML中，第1000-1008页，2014年。

肖克里和维塔利·什马蒂科夫。隐私保护的深度学习。在第22届NdACM SIGSAC计算机和通信安全会议会议记录中，CCS‘15,2015。

speedtest.net.速度最快的市场回购协议rt.http://www.speedtest.net/reports/united-states/，2016年8月。

约斯特·托拜厄斯·斯普林恩伯格、阿列克谢·多索维茨基、托马斯·布罗克斯和马丁·里德米勒。力求简单：全卷积网络。arXiv:1412.6806, 2014.

苏雷什，费利克斯x余，桑吉夫库马尔，和H.布伦丹麦克马汉。具有有限通信的分布式平均估计。第34届机器学习国际会议论文集，第3页。3329–3337, 2017.

大卫·p·伍德拉夫。草图作为数值李近代数的工具。理论计算机科学的基础和趋势，10（12）：1-157,2014年。ISSN1551-305X.doi: 10.1561/0400000060.

张宇晨和小林。迪斯科：分布优化离子的自我一致的经验损失。在ICML，pp。362–370, 2015.