

LEAF：联邦设置的基准



卡巴斯蒂安，卡迪安，∗，11112

H、 布伦丹·麦克马汉，弗吉尼亚·史密斯，阿梅特·塔尔沃卡211*,*3

1 卡内基梅隆大学

2 谷歌

3 确定的人工智能https://leaf.cmu.edu

# 摘要

现代的联邦网络，例如由可穿戴设备、移动电话或自动驾驶车辆组成的网络，每天都会产生大量的数据。这些丰富的数据有助于学习模型，这些模型可以改善每个设备上的用户体验。然而，联邦数据的规模和异构性给联合学习、元学习和多任务学习等研究领域带来了新的挑战。随着机器学习社区开始应对这些挑战，我们正处于一个关键时刻，以确保在这些领域取得的进展以现实的基准为基础。为此，我们提出了LEAF，一个用于在联邦环境中学习的模块化基准测试框架。LEAF包括一套开源的联邦数据集、严格的评估框架和一组参考实现，所有这些都是为了捕捉实际联邦环境中的障碍和复杂性。

# 1        介绍

随着越来越多的数据在远程设备的联合网络上生成，人们越来越感兴趣的是使用这些数据的模型来授权设备上的应用程序[22,23,30,19,37]。然而，学习联邦网络中生成的数据会带来几个新的障碍：

统计：在每个设备上以一种异构的方式生成数据，每个设备与一个不同的（尽管可能相关）底层数据生成分布相关联。此外，数据点的数量通常在不同的设备上有很大的不同。

系统：联邦场景中的设备数量通常比典型的分布式环境（如数据中心计算）中的节点数量大一个数量级。此外，每个设备在存储、计算和通信容量方面可能具有显著的限制。此外，由于硬件、网络连接和功率的变化，这些容量也可能因设备而异。因此，联邦设置可能会遇到通信瓶颈，这使传统数据中心设置中遇到的瓶颈相形见绌，并且可能需要更快的设备内推理。

隐私和安全性：最后，个人生成数据的敏感特性要求对联合数据进行操作的方法，以平衡隐私和安全问题与更传统的考虑因素，如统计准确性、可伸缩性和效率[24,4]。

最近的研究提出了不同的方法来应对这些挑战，但是当涉及到它们的实验评估时，这些努力中的许多是不够的。作为一个例子，考虑联邦



∗

通讯作者。通信地址：斯皮尔达斯@cmu.edu。

第33届神经信息处理系统会议（NeurIPS 2019），加拿大温哥华。

学习范式，集中于直接在联邦网络上训练模型[22，30，28]。以联合学习为重点的实验工作广泛地使用了三种类型的数据集，每种类型都有各自的缺点：（1）常用的数据集不能提供联邦场景的真实模型，例如MNIST、MNIST方式或CIFAR-10的人工分区[22,13,8,3,11,32,34]；（2） 真实但专有的联合数据集，例如[22]中来自未命名社交网络的数据，[18]中的众包语音命令，[5]中的华为专有数据；以及（3）真实的联邦数据集，这些数据源于公共可用数据，但不容易复制，例如[25]中的FaceScrub，莎士比亚的[22]和[13，24，3]中的Reddit。

作为第二个例子，考虑元学习，一种由[5]和[12]提出的相关学习范式，作为解决联邦网络统计挑战的一种方法。这种模式确实很适合联邦环境，因为异构设备可以解释为元学习任务。然而，流行的元学习基准，如Omniglot[15，7，33，31]和miniImageNet[29，7，33，31]关注的是快照学习（即，所有任务都有相同数量的样本，每个班级在每个任务中有相同数量的样本，因此未能捕捉到联邦数据给元学习解决方案带来的现实挑战。事实上，前面提到的所有数据集都可以归为上面提到的第一种类型（在我们的目的中流行但不现实）。*k*

最后，考虑多任务学习（MTL）。这种模式也适用于联邦环境[30]，但与现实的联邦网络相反，通常在任务和样本数量较少的制度中进行探索，例如，流行的地雷探测[38,26,36,30]，计算机调查[2,1,14]和伦敦教育局内学校[26,17,1,2，14] 每个数据集最多有200个任务。我们强调，虽然联合学习、元学习和多任务学习是LEAF的应用，但该框架实际上包含了广泛的潜在学习设置，例如设备上学习或推理、转移学习、终身学习和个性化学习模型的开发。

我们的工作旨在弥合人工数据集之间的差距，人工数据集是受欢迎的，可用于基准测试，与那些真实地捕捉联邦场景特征的数据集之间的差距，但到目前为止，这些数据集要么是专有的，要么是难以处理的。此外，除了建立一套联邦数据集之外，我们还提出了评估方法和重现结果的清晰方法。为此，我们提出了LEAF，一个模块化的基准测试框架，用于远程设备的大规模分布式联邦网络中的学习。

# 2        叶子

LEAF是联邦设置的开源基准。它包括（1）一套开源数据集，（2）一系列统计和系统指标，以及（3）一套参考实现。如图1所示，LEAF的模块化设计允许这三个组件轻松地集成到不同的实验管道中。我们继续详细介绍LEAF的核心组件。[[1]](" \l "_ftn1" \o ")

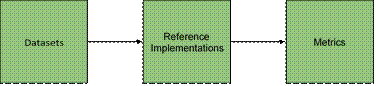
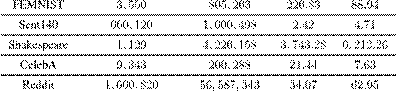


图1：叶模块。“数据集”模块对数据进行预处理并将其转换为标准格式，该格式可以集成到任意ML管道中。LEAF的“参考实现”模块是一个不断增长的存储库，其中包含了联邦设置中使用的常用方法，每个实现都会生成一个包含各种不同统计和系统度量的日志。任何以适当格式生成的日志都可以通过LEAF的“metrics”模块以各种方式聚合和分析这些度量。

表1:LEAF中的数据集统计。







数据集：我们为LEAF策划了一套真实的联邦数据集。我们关注的是数据集，其中（1）数据具有自然的键控生成过程（其中每个键都指向特定的设备/用户）；（2）数据是从数千到数百万个设备的网络中生成的；以及（3）数据点的数量在设备之间倾斜。目前，LEAF包含六个数据集：

•，它是根据数字/字符的写入者在扩展MNIST[16，6]中对数据进行分区而构建的。*联邦扩展MNIST（FEMNIST）*

•[9]，一个自动生成的情绪分析数据集，根据推文中的表情符号对其进行注释。每个设备都是不同的twitter用户。*感伤140*

•《莎士比亚全集》中的数据集[35,22]。在这里，每出戏中的每个说话角色都被视为一种不同的手段。*莎士比亚*

•，根据照片上的名人对大规模的名人空间属性数据集[21]进行划分。*塞莱巴*[[2]](" \l "_ftn2" \o ")

•，我们对2017年12月在社交网络上发布的评论进行了预处理。*雷迪特*

•一个合成数据集，它修改了[20]中给出的合成数据集，使其对当前的元学习方法更具挑战性。详见附录A。

我们在表1中提供了这些数据集的统计数据（除了合成数据集，因为这些数据集因用户的设置而异）。在LEAF中，我们为每个数据集提供所有必要的预处理脚本，以及原型和最终测试的小型/完整版本。接下来，我们计划添加来自不同领域（如音频、视频）的数据集，并增加机器学习任务的范围（例如文本到语音、翻译、压缩等）。

度量：需要严格的评估度量来适当地评估学习解决方案在联合场景中的行为。目前，LEAF建立了一组专门为此目的而选择的度量标准。例如，我们引入了更好地捕捉整个设备性能分布的指标：第10、第50和第90百分位的性能以及按数据中自然层次结构分层的性能（例如，莎士比亚数据集中的“play”或Reddit的“subreddit”）。我们还介绍了一些指标，这些指标根据flop数和下载/上传的字节数来计算边缘设备所需的计算资源量。最后，LEAF还认识到了指定如何在设备之间加权精度的重要性，例如，是每个设备都同等重要，还是每个数据点都同等重要（意味着高级用户/设备获得优惠待遇）。

参考实现：为了促进可再现性，LEAF还包含一组面向联邦场景的算法参考实现。目前，这个集合仅限于联邦学习范式，尤其包括minibatch SGD、FedAvg[22]和Mocha[30]的参考实现。我们的目标是在更广泛的研究社区的帮助下，为LEAF提供更多的方法和范例的实现。

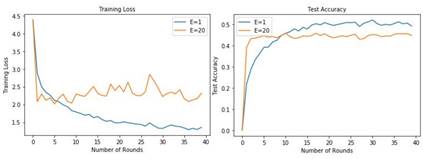


图2:FedAvg在莎士比亚数据集的子样本上的收敛行为。我们能够获得与[22]中获得的结果相当的每个样本测试精度。我们还定性地复制了在大量局部时期观察到的训练损失差异（）。*E*

# 3        叶在行动

现在我们来看看叶在行动。我们特别强调LEAF的三个特点：[[3]](" \l "_ftn3" \o ")

LEAF支持可复制的科学：为了证明通过LEAF实现的可再现性，我们专注于定性地再现[22]在莎士比亚数据集上获得的结果，以用于下一个字符预测任务。特别值得注意的是，对于这个特定的数据集，FedAvg方法会随着局部时间段的数量增加而出人意料地发生变化。因此，在部署FedAvg等方法之前，这是一个需要了解的关键设置。结果如图2所示，在图2中，我们确实看到了类似的发散行为，在训练损失方面，我们增加了纪元的数量。

在图3中，我们建议将多个统计指标作为重要指标同时提供给客户。对于统计指标，我们展示了在感伤140（我们表示为）中改变每个用户的最小样本数的效果。我们看到，虽然数据不足的用户的平均性能只会略有下降（即=3），但第25个百分位则会急剧下降。同时，对于系统指标，我们为FEMNIST运行minibatch SGD和FedAvg，并计算达到每个样本精度阈值0.75所需的系统预算。我们以所有设备的总浮点数和上传到网络的总字节数来描述预算。我们的结果表明，当涉及到通信与本地计算的权衡时，FedAvg的系统轮廓得到了改善，尽管我们注意到，在这两个维度上，一般的方法可能有所不同。*kk*

LEAF是模块化的：为了证明LEAF的模块化，我们将它的“Datasets”模块合并到三个新的实验管道中：一个为每个设备训练纯本地模型（在CelebA和我们的合成数据集上），一个忽略设备之间的自然分区，即混合所有数据（在Reddit上），以及一个使用流行的元学习方法爬行动物[27]（在女性身上）。这些实验的结果如表2所示。这些特殊的管道揭示了不同的建模方法对不同的联邦数据集是多么的合适[10，12]。

# 4        结论

我们介绍了LEAF，一个在联邦环境下学习的模块化框架，或者以大规模分布式设备网络为标志的生态系统。在这种环境下适用的学习范式包括联合学习、元学习、多任务学习和设备上学习。

LEAF将允许联合学习、元学习和多任务学习等领域的研究人员和实践者在更现实的假设下对新提出的解决方案进行推理

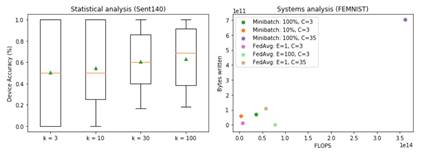


图3:Sent140和FENIST的统计和系统分析。对于Sent140：是每个用户的最小样本数。橙色线表示设备的中值精度，绿色三角形表示平均值，方框表示第25和75个百分位，胡须代表第10到90个百分位。对于FEMNIST：是每轮选择的客户数量，是每个客户在本地接受FedAvg培训的时期数。对于minibatch SGD，我们报告每个客户端使用的数据百分比。*kCE*

表2:LEAF的模块化演示。我们将LEAF的数据集整合到新的实验管道中（FedAvg之外），并报告由此产生的样本测试精度。



                                      数据集FedAvg（基线）附加管道

|  |  |  |
| --- | --- | --- |
|  |  | 描述准确性 |

                                      塞莱巴89.46%65.29%

本地模型

|  |  |  |  |
| --- | --- | --- | --- |
| 合成 | 71.89% |  | 87.34% |
| 雷迪特 | 13.35% | 全球IID模型 | 12.60% |
| 女性主义者 | 74.72% | 爬行动物 | 80.24% |

以前的基准。我们打算让LEAF了解最新的数据集、指标和开源解决方案，以促进在这一领域的知情和有根据的进展。

致谢

这项工作得到了美国国防部高级研究计划局FA875017C0141的部分支持，这是国家科学基金会的资助

IIS1705121和IIS1838017，Okawa拨款，Google教师奖，Amazon网络服务

该奖项是摩根大通a.I.研究人员奖、卡内基博世研究所研究奖，以及康尼克斯研究中心，JUMP六个中心之一，DARPA赞助的半导体研究公司（SRC）项目。本材料中表达的任何意见、发现和结论或建议均为作者的观点，并不一定反映DARPA、国家科学基金会或任何其他资助机构的观点。

# 工具书类

[1] 阿文·阿加瓦尔、塞缪尔·格伯和哈尔·道姆。使用流形正则化学习多个任务。神经信息处理系统进展，2010。

[2] 安德烈亚斯·阿格里欧、西奥多罗斯·叶夫根尼奥和马西米利亚诺·庞蒂尔。凸多任务特征学习。机器学习，73（3）：243–2722008年。

[3] 安德烈·达维塔斯、安德烈·达维塔斯、德布什维塔斯、德布什维塔斯、德布什维塔斯、巴格什维塔斯。如何后门联合学习。arXiv预印本arXiv:1807.004592018年。

[4] 基思·博纳维茨、弗拉基米尔·伊万诺夫、本·克鲁特、安东尼奥·马塞东、H·布伦丹·麦克马汉、萨瓦尔·帕特尔、丹尼尔·拉马奇、亚伦·西格尔和卡恩·塞思。保护隐私机器学习的实用安全聚合。2017年ACM SIGSAC计算机和通信安全会议。

[5] 陈飞，董振华，李振国，何秀强。联合元学习推荐。arXiv预印本arXiv:1802.078762018年。

[6] 格雷戈里·科恩、赛义德·阿夫沙、乔纳森·塔普森和安德烈·范沙克。EMNIST:MNIST到手写字母的扩展。arXiv预印本arXiv:1702.053732017年。

[7] 切尔西·芬恩，彼得·阿贝尔和谢尔盖·莱文。模型不可知元学习快速适应深层网络。2017年机器学习国际会议。

[8] Robin C Geyer，Tassilo Klein和Moin Nabi。差异私人联合学习：客户层面的视角。arXiv预印本arXiv:1712.075572017年。

[9] Alec Go，Richa Bhayani和Lei Huang。基于远程监控的Twitter情感分类。项目报告，斯坦福大学，2009年。

[10] 江宜汉、雅库布·科尼肯、基思·拉什和斯雷拉姆·坎南。通过模型不可知元学习改进联合学习的个性化。arXiv预印本arXiv:1909.124882019年。

[11] 迈克尔·坎普、利娜拉·阿迪洛娃、约阿希姆·希奇、法比安·Hüger、彼得·施利希特、蒂姆·维尔茨和斯特凡·沃贝尔。基于动态模型平均的有效分散深度学习。在2018年欧洲数据库机器学习和知识发现联合会议上。

[12] 米哈伊尔·霍达克、玛丽亚·弗洛里娜·巴尔坎和阿梅特·塔尔沃卡尔。基于自适应梯度的金属学习方法。神经信息处理系统进展，2019年。

[13] Jakub Konecnˇy、H Brendan McMahan、Felix X Yu、Peter Richtárik、Ananda Theertha Suresh和Dave Bacon。联合学习：提高沟通效率的策略。arXiv预印本arXiv:1610.054922016年。

[14] Abhishek Kumar和Hal Daume III.多任务学习中的学习任务分组和重叠。arXiv预印本arXiv:1206.64172012年。

[15] 布伦登湖，Ruslan Salakhutdinov，Jason Gross和Joshua Tenenbaum。简单视觉概念的一次性学习。在2011年认知科学学会年会上。

[16] 杨乐坤。手写数字的MNIST数据库。http://yann.lecun.com/exdb/mnist/，1998年。

[17] 李纪勋，杨恩浩和黄星焕。基于任务相关性和损失的非对称多任务学习。2016年国际机器学习大会。

[18] 大卫·勒罗伊、爱丽丝·库克、蒂鲍特·拉夫里尔、蒂鲍特·吉斯尔布雷希特和约瑟夫·杜罗。关键字识别的联合学习。IEEE国际声学、语音和信号处理会议，2019年。

[19] 田莉，阿尼特·库马尔·萨胡，阿梅特·塔尔瓦卡和维吉尼亚·史密斯。联合学习：挑战、方法和未来方向。arXiv预印本arXiv:1908.078732019年。

[20] 田莉、马齐亚尔·桑贾比和维吉尼亚·史密斯。公平学习资源分配。arXiv预印本arXiv:1905.104972019年。

[21]刘紫薇，罗萍，王晓刚，汤晓鸥。深入学习野外面孔属性。2015年国际计算机视觉会议。

[22]H布伦丹·麦克马汉、艾德·摩尔、丹尼尔·拉马奇、塞思·汉普森和布莱斯·阿奎拉·阿卡斯。从分散数据中有效地学习深层网络。人工智能与统计，2017年。

[23]H布伦丹·麦克马汉和丹尼尔·拉马奇。联合学习：没有集中训练数据的协作机器学习。googlelogs.com网站2017年。

[24]H Brendan McMahan、Daniel Ramage、Kunal Talwar和Li Zhang。学习差异私人递归语言模型。2018年国际学习代表大会。

[25]卢卡·梅利斯、宋聪正、埃米利亚诺·德克里斯托法罗和维塔利·什马蒂科夫。针对协作学习的推理攻击。arXiv预印本arXiv:1805.040492018年。

[26]Keerthiram Murugesan和Jaime Carbonell。多任务多核关系学习。2017年暹罗国际数据挖掘会议。

[27]亚历克斯·尼科尔、约书亚·阿奇姆和约翰·舒尔曼。关于一阶元学习算法。arXiv预印本arXiv:1803.029992018年。

[28]Vasyl Pihur、Aleksandra Korolova、Frederick Liu、Subhash Sankuratipati、Moti Yung、Dachuan Huang和Ruogu Zeng。差分私有“取舍”机器学习。arXiv预印本arXiv:1807.043692018年。

[29]萨钦·拉维和雨果·拉罗谢尔。优化作为少镜头学习的模型。https://openreview.net/pdf？id=rJY0 Kcll，2016年。

[30]维吉尼亚·史密斯、蒋超楷、马齐亚尔·桑贾比和阿梅特·S·塔尔瓦卡。联合多任务学习。神经信息处理系统进展，2017年。

[31]杰克·斯内尔、凯文·斯沃斯基和理查德·泽梅尔。用于少镜头学习的典型网络。神经信息处理系统进展，2017年。

[32]格雷戈·乌尔姆、埃米尔·古斯塔夫松和马茨·吉尔斯特朗。erlang中的函数联合学习（ffl-erl）。2018年国际功能与约束逻辑编程研讨会。

[33]奥利奥·温尼尔斯、查尔斯·布伦德尔、蒂姆·利利克拉普、达安·维尔斯特拉等。匹配网络进行一次性学习。神经信息处理系统进展，2016年。

[34]王世强，蒂凡尼·图尔，西奥多罗斯·萨洛尼迪斯，梁建基，马卡娅，何婷，陈凯文。资源受限边缘计算系统中的自适应联合学习。IEEE通信选定领域期刊，2019年。

[35]威廉·莎士比亚。莎士比亚全集。可在//www.gutenberg.org/ebooks/100。

[36]雅雪，廖学军，劳伦斯·卡林，巴拉吉·克里希纳普拉姆。基于dirichlet过程先验的多任务分类学习。机器学习研究杂志，8（1月）：35-632007。

[37]杨强，刘洋，陈天健，童永新。联合机器学习：概念与应用。ACM智能系统与技术交易，2019年。

[38]张毅和杰夫G施耐德。用稀疏矩阵法向惩罚学习多个任务。神经信息处理系统进展，2010。

# 合成数据集

我们的合成数据集（在第2节中介绍）的灵感来自于[20]中提出的数据集，但它可能具有额外的异构性，旨在使当前的元学习方法（如爬行动物[27]）失败。高级目标是创建真正的模型（1）依赖于任务的任务，以及（2）聚集在不止一个中心的任务。

首先，用户必须输入所需的设备数量≥1和一个向量（p1，…，pk）*T*就这样。作为生成任务的准备：

1.    样本聚类意味着∈R∈1，…，k。为此，绘制∼N（Bj，I），Bj∼N（0，I）。*µjs，jµj*

2.    绘制矩阵∈Rby抽样∼N（0，I）。*问d*+1×秒*问*

3.    创建对角线矩阵∑，使∑i，i=i−1.2。

现在，对于每个任务∈1，…，T：*t*

1.    根据输入概率（p1，…，pk）对聚类中心进行抽样。*微特*

2.    绘制∼N（µt，I）并设置为Qut，wt R。*美国犹他州重量*∈*d*+1

3.    现在，从平均值为3和σ2的对数正态分布中提取。然后我们将samples的数量设置为min（mt+51000）（为每个任务的样本数设置一个下限和一个上限）。*公吨新台币*

4.    样品∼N（Ct，I），Ct∼N（0，I）。*及物动词*

5.    现在，对于∈1，…，nt*我*，抽样图。

6.    最后，在添加必要的填充以解释截距后，设置=argmax（sigmoid（wtxit+N（0,0.1·I）））。*伊提xit公司*

# B实验细节

在本节中，我们将详细介绍第3节中介绍的实验。

莎士比亚汇合。对于图2中所示的实验，我们对118个设备进行了子抽样

（约占总数的5%）在我们莎士比亚的资料中。我们的模型首先将每个字符映射到维度8的嵌入，然后将其传递到一个由两层256个单元组成的LSTM。LSTM发出一个输出嵌入，通过点积和一个softmax对词汇表的所有项目进行评分。我们对LSTM使用80的序列长度。我们使用AccuracyTop1进行评估。我们使用每轮0.8和10个设备的学习率进行所有实验。

统计和系统分析。对于图3中所示的所有Sent140实验，我们使用了带逻辑回归的单词袋模型，学习率为3·10−4。对于同一张图中的FEMNIST实验，我们对5%的数据进行了亚采样，并使用了一个模型，该模型有两个卷积层和一个最终密度层（2048个单元）。FedAvg的学习率为4·10−3，小批量SGD的学习率为6·10−2。

附加管道。对于表2中所示的实验，我们对每个用户使用60%的培训、20%的验证和20%的测试，并在测试集上报告结果。每个实验的超参数如下：

•对于CelebA实验，我们使用了10%的客户和上面描述的FEMNIST相同的模型。对于本地模型，每个设备在[0.1,0.01,0.001,0.0001]中探索了学习率。FedAvg模型每轮使用10个客户机进行100轮训练，本地训练一个时代，批量大小为5，最佳学习率为0.001。两个结果在5次运行中的平均值。

•对于合成数据集的实验，我们使用1000台设备，仅使用一个群集、60个特征和5个类。我们的模型是一个乙状结肠激活的感知器。对于本地模型，每个设备在[10−3,10−2,10−1,1,10102103]中探索了学习率。FedAvg模型每轮使用10个客户，每次100轮，在本地培训一个时期，批大小为5，发现最佳学习率为0.1。

•对于Reddit实验，我们使用了819个设备和一个与我们为莎士比亚描述的模型相似的模型。主要的区别是：嵌入的大小现在是200，我们从训练集中的令牌构建词汇表，固定长度为10000。我们使用10的序列长度，使用AccuracyTop1进行计算，并认为所有未知和填充标记的预测都是不正确的。对于全球iid模型，我们在所有设备上训练了3个时代√



数据学习率为4.2。对于FedAvg，我们每轮使用10个客户机进行100轮训练，在本地使用5个批次进行一个时代的培训。我们使用8的学习率。两个结果在5次运行中的平均值。

•对于FEMNIST实验，我们使用前面描述的相同模型，并运行每个算法1000轮，每轮使用5个客户机，本地学习率为10-3，5个小批量的训练小批量为10，并在一组看不见的测试设备上进行评估。此外，对于爬行动物，我们使用从2到0的线性衰减的元学习率，并通过对每个测试设备进行微调来评估大小为5的50个小批次。

[[1]](" \l "_ftnref1" \o ") 所有代码和文档都可以在https://github.com/TalwalkarLab/leaf/。

[[2]](" \l "_ftnref2" \o ") 原始CelebA数据托管在http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

[[3]](" \l "_ftnref3" \o ") 实验细节见附录B。