联邦学习移动键盘预测

*安德鲁·哈德、Kanishka Rao、Rajiv Mathews、Swaroop Ramaswamy、Franc¸oise Beaufays*

*肖恩·奥根斯坦，休伯特·艾希纳，克洛伊·基登，丹尼尔·拉马奇’*

谷歌有限责任公司，

美国加利福尼亚州山景城。

{哈达，卡尼什卡劳，马修斯，斯沃鲁普兰，fsb索根斯特，胡伯特，洛基，戏剧}@谷歌

# 摘要

为了在智能手机虚拟键盘上进行下一个单词预测，我们使用一个称为联邦学习的分布式设备学习框架来训练递归神经网络语言模型。将随机梯度下降的基于服务器的训练与使用联邦平均算法的客户端设备上的训练进行了比较。联邦算法可以在这个用例的更高质量的数据集上进行训练，它可以实现更好的预测召回。这项工作证明了在客户端设备上训练语言模型的可行性和好处，而无需将敏感的用户数据导出到服务器。联邦学习环境使用户能够更好地控制其数据的使用，并简化了默认情况下通过分布式培训和聚合合并隐私的任务

客户端设备的填充。

*索引术语*-联合学习，键盘，语言建模，NLP，CIFG。

# 1.    简介

谷歌键盘[1]是一款用于触摸屏移动设备的虚拟键盘，截至2019年，它支持600多种语言，安装超过10亿次。除了解码来自输入模式的噪声信号（包括轻敲和单词手势输入），Gboard还提供自动校正、单词完成和下一个单词预测功能。

随着用户越来越多地转向移动设备[1]，可靠和快速的移动输入方法变得越来越重要。下一个单词的预测提供了一个方便文本输入的工具。语言模型（LMs）基于少量的用户生成的前文本，可以预测最可能的下一个单词或短语。图1提供了一个示例：给定文本“我爱你”，Gboard预测用户接下来可能会键入“and”、“too”或“so much”。建议条中的中心位置保留为最大概率

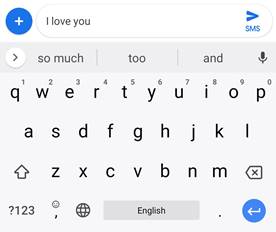


图1。Gboard中的下一个单词预测。根据“我爱你”的上下文，键盘预测“和”、“太”和“太多了”。

候选人，而第二位和第三位最有可能的候选人分别占据左右位置。

在这项工作之前，预测是用一个单词n-gram有限状态传感器（FST）[2]生成的。Gboard中FST解码器的机制-包括FST在字面解码、校正和补全中的作用在参考文献[3]中描述。下一个单词的预测是通过搜索与前面的文本相匹配的最高阶n-gram状态来构建的。将返回此状态的-最佳输出标签。还考虑了包含到较低阶的退避转换的路径。Gboard中英语语言的主要（静态）语言模型是Katz平滑贝叶斯插值[4]5-gram LM，包含125万个n-gram，包括164000个unigram。个性化的用户历史记录、联系人和电子邮件n-gram模型增强了主LM。*n*

移动键盘模型受到多种方式的约束。为了同时在低端和高端设备上运行，模型应该很小，推理时间延迟应该是

低一点。用户通常希望在输入事件发生后20毫秒内出现可见的键盘响应。考虑到移动键盘应用程序的使用频率，如果不限制CPU消耗，客户端设备的电池可能会很快耗尽。因此，语言模型的大小通常被限制在几十兆字节的范围内，词汇表有几十万个单词。

神经模型——特别是单词和字符级的递归神经网络（RNNs）[5]——已经被证明在语言建模任务上表现良好[6，7，8]。与RNNs不同的是，RNNs和前向神经网络使用的是基于固定上下文的动态窗口。时间反向传播算法中的爆炸和消失梯度可以用长短期记忆（LSTM）来解决[6]。在撰写本文时，10亿字基准测试[9]的最新技术难题已经通过LSTM变体[10，11]实现。

训练预测模型需要一个大的数据样本来代表用户将提交的文本。可以使用公共可用的数据集，尽管训练分布通常与总体分布不匹配。另一个选择是对用户生成的文本进行采样。这需要日志记录、基础设施、服务器上的专用存储和安全性。即使有了数据清理协议和严格的访问控制，用户也可能对其个人数据的收集和远程存储感到不安[12]。

在本文中，我们展示了在商业环境中，联合学习提供了一种替代基于服务器的数据收集和训练范例的方法。我们在服务器和联邦环境中从头开始训练RNN模型，并在FST解码器基线方面实现召回改进。

论文按以下方式组织。第2节总结了与移动输入解码、RNNs语言建模和联合学习相关的前期工作。耦合输入遗忘门（CIFG）-用于下一个字预测的RNN变体-在第3节中描述。第4节

更深入地讨论了联邦平均算法。总结了基于server5的联邦模型训练和实验。研究结果见第6节，结束语见第7节。

# 2.    相关工作

fst已经在移动键盘输入解码、校正和预测的背景下进行了研究[3]。LSTMs极大地改进了移动键盘上手势输入的解码[13]。在推理时间延迟和内存限制下，为单词预测率和键击节省而优化的RNN语言模型也已经发表[14,15]。

神经网络模型的分布式训练研究

随着最近对隐私和政府监管的日益重视，这一点得到了重视。特别是，联合学习已经被证明是基于服务器的分布式训练到使用本地存储数据的基于客户端设备的训练的有用扩展[12,16]。语言模型已经使用联邦算法和差分隐私相结合进行了训练[17,18]。Gboard以前曾使用联合学习来训练一个模型，以根据输入上下文[19]来建议搜索查询，不过结果还没有公布。据我们所知，目前还没有出版物为使用联合学习的移动键盘训练神经语言模型。

# 3.    模型体系结构

下一个单词预测模型使用长短期记忆（LSTM）[6]递归神经网络的变体，称为耦合输入遗忘门（CIFG）[20]。与选通递归单元[21]一样，CIFG使用单个门来控制输入和循环单元的自连接，从而将每个单元的参数数量减少了25%。对于timestep，输入门和遗忘门具有以下关系：*t它英尺*

*英尺*=1−i.t（一）

CIFG体系结构在不影响模型性能的前提下，减少了计算量和参数集大小，有利于移动设备环境。模型使用张量流[22]进行训练，无需窥视孔连接。TensorFlow Lite[2]支持设备内推理。

结合输入嵌入和输出投影矩阵来减小模型大小和加快训练速度[23,24]。在给定的词汇量下，一个热编码∈Ris映射到一个稠密嵌入向量∈Rby=Wv，嵌入矩阵∈R，CIFG的输出投影映射到输出向量∈R，输出向量上的softmax函数将原始逻辑转化为规范化概率。输出和目标标签的交叉熵损失用于训练。*五五五dDdWD*×V型*DW*T*h五*

第1节中提到的客户端设备要求限制了词汇表和模型大小。输入和输出词汇表使用=10000个单词的字典。输入标记包括特殊的句首、句尾和词汇外标记。在网络评估和推理过程中，与这些特殊标记相对应的logit被忽略。输入嵌入和CIFG输出投影维度设置为96。使用了670个单元的单层CIFG。总的来说，140万个参数组成了这个网络，其中超过三分之二与嵌入矩阵有关。经过权重量化后，发送到Gboard设备的模型大小为1.4兆字节。*五DW*

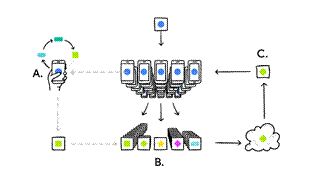


图2。参考文献[19]中联邦学习过程的一个例子：（A）客户端设备根据本地存储的数据计算SGD更新，（B）服务器聚合客户端更新以构建新的全局模型，（C）将新模型发送回客户端，并重复该过程。

# 4.    联合学习

联邦学习[12，16]提供了一种分散计算策略，可以用来训练神经模型。移动设备被称为客户机，产生大量的个人数据，可用于培训。客户机处理本地数据并与服务器共享模型更新，而不是将数据上载到服务器进行集中培训。来自大量客户端的权重由服务器聚合并组合以创建改进的全局模型。图2展示了这个过程。分布式方法已经被证明可以处理不平衡的数据集和不独立或不完全分布在客户机上的数据。

在服务器上使用联邦平均算法[12]，以组合客户端更新并生成新的全局模型。在训练轮，一个全局模型被发送到客户端设备的子集。在=0的特殊情况下，客户机设备从相同的全局模型开始，该模型要么是随机初始化的，要么是在代理数据上预先训练的。参与给定回合的每个客户机都有一个由示例组成的本地数据集，其中是参与客户机的索引。因设备而异。对于Gboard的研究，是与用户的打字量有关的。*t重量Ktnk公司knk公司nk公司*

每个客户机根据当前模型的本地数据计算平均梯度*吉克重量*使用一个或多个随机梯度下降（SGD）步骤。对于客户端学习率，本地客户端更新由以下公式给出：

*.* （二）

然后，服务器对客户机模型进行加权聚合，以获得新的全局模型：*重量*+1

*,* （三）

其中=。本质上，客户机在本地计算SGD更新，这些更新被传递到服务器并进行聚合。超参数包括客户机批处理大小、客户机时间段数和每轮客户机数量（全局批处理大小）都经过调整以提高性能。*N*P*knk公司*

分散的设备上计算比服务器存储提供更少的安全和隐私风险，即使服务器托管的数据是匿名的。将个人数据保存在客户端设备上可以让用户对自己的数据进行更直接和更实际的控制。每个客户机传递给服务器的模型更新是短暂的、集中的和聚合的。客户机更新从不存储在服务器上；更新在内存中处理，并在累积到权重向量后立即丢弃。遵循数据最小化的原则[25]，上传内容仅限于模型权重。最后，结果仅用于聚合：通过合并来自多个客户端设备的更新来改进全局模型。这里讨论的联合学习过程要求用户相信聚合服务器不会仔细检查单个权重上传。这仍然比服务器培训更可取，因为服务器从来没有委托给用户数据。正在探索其他技术来放宽信任要求。联合学习已经被证明是对隐私保护技术的补充，如安全聚合[26]和差异隐私[17]。

# 5.    实验

联邦学习和基于服务器的随机梯度下降用于从随机权值初始化开始训练第3节中描述的CIFG语言模型。这两个模型的性能都是在服务器托管日志数据、客户端保存的数据和实际生产实验中评估的。

5.1条。基于服务器的日志数据训练

基于服务器的CIFG下一个单词预测模型的培训依赖于Gboard用户记录的数据，这些用户在输入googleapps时选择共享文本片段。文本被截断以包含几个单词的短短语，而片段只是偶尔从单个用户记录下来的。在培训之前，日志是匿名的，并删除个人身份信息。此外，如果片段以句首标记开头，则仅用于训练。

在这项研究中，日志是从美国讲英语的Gboard用户中收集的。大约有75亿句话用于培训，而测试和评估样本各包含25000句。数据集中的平均句子长度为4.1个单词。表1提供了按应用程序类型划分的日志数据。聊天应用程序生成大部分记录文本。

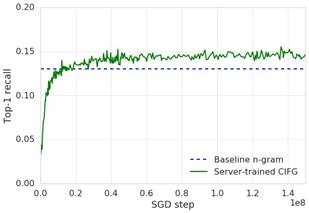


图3。在服务器培训期间，将CIFG作为SGD步骤的一个功能进行Top-1调用。为了比较，我们展示了n-gram FST基线模型的召回率，但是在本研究中没有训练FST模型。

异步随机梯度下降的学习率等于10−3，没有重量衰减或动量衰减用于训练服务器CIFG。自适应梯度法，包括Adam[27]和AdaGrad[28]都没有发现改善收敛性的方法。句子以50个为一批进行处理。网络经过1.5亿新元的步进后融合。图3显示了在网络训练期间，与n-gram基线模型的性能相比，CIFG的前1召回率。

|  |  |
| --- | --- |
| 应用程序类型 | 数据共享 |
| 聊天 | 60% |
| Web输入 | 35% |
| 长格式文本 | 5% |

表1。移动应用程序类型的日志数据组成。

5.2条。使用客户端缓存的联合训练

用于CIFG下一个字预测模型的联合训练的数据存储在Gboard客户端设备上的本地缓存中。与日志数据一样，每个客户端缓存存储属于设备所有者的文本，以及解码器生成的预测候选。

客户端设备必须满足许多要求，才能有资格参加联合培训。就硬件要求而言，这些设备必须至少有2GB的可用内存。此外，只有在客户机正在充电、连接到未计量网络且处于空闲状态时，才允许参与。这些标准是专门为联邦学习的Gboard实现而选择的，并不是联邦学习平台固有的。本研究的客户也必须位于

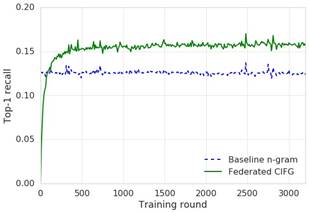


图4。在联合训练期间，作为训练轮次的一项功能，对CIFG进行了Top-1召回。ngram-FST基线模型的性能在客户端缓存和CIFG上进行了评估，但在本研究中没有对其进行培训。

北美，同时运行Gboard 7.3版或更高版本，并启用美国英语语言模型。

不同于通过训练和测试获得的独立样本，通过训练和测试获得独立的样本。虽然没有明确地将客户机设备分为三个不同的群体，但是在足够大的客户机群体中，客户机在培训和测试或评估任务中重用的概率是最小的。表2显示了按应用程序类型划分的客户端缓存数据的组成。与日志数据一样，客户端缓存也由聊天应用程序控制。社交媒体应用程序在客户端缓存样本中的出现率有所增加，而长形式通信则较少。

|  |  |
| --- | --- |
| 应用程序类型 | 数据共享 |
| 聊天 | 66% |
| 社交 | 16% |
| Web输入 | 5% |
| 其他 | 12% |

表2。按移动应用程序类型划分的客户端缓存数据的组成。

第4节中描述的联邦平均算法用于聚合分布式客户端SGD更新。在Gboard中结束每一轮联合培训需要100到500个客户端更新。等式3中的服务器更新通过动量优化器实现，使用Nesterov加速梯度[29]，动量超参数为0.9，服务器学习率为1.0。这项技术被发现可以减少训练时间的替代品，包括纯SGD。平均来说，每个客户在一个培训周期内处理大约400个例句。联邦CIFG在经过3000轮培训后汇聚在一起，在这个过程中，150万客户处理了6亿个句子。培训通常需要4-5天。联邦CIFG的前1个召回在图4中显示为培训轮次的函数。在联邦eval任务中也对n-gram基线模型的性能进行了测量，以便为CIFG提供一个比较，尽管解码器在本研究中没有经过训练。Ngram模型召回是通过比较设备上训练缓存中的候选解码器与实际用户输入的文本来衡量的。

# 6.    结果

每个模型的性能都是使用召回指标来评估的，该指标定义为正确预测的数量与令牌总数的比率。对于Gboard来说，召回最有可能的候选词是很重要的，因为用户更倾向于阅读和利用中心建议点的预测。由于Gboard在建议栏中包括三名候选人，前三名的召回也值得关注。

|  |  |  |
| --- | --- | --- |
| 模型 | 前1名召回 | 前三名召回 |
| 纳克 | 13.0% | 22.1% |
| 服务器CIFG | 16.5% | 27.1% |
| 联合CIFG | 16.4% | 27.0% |

表3。服务器和联合CIFG模型的预测召回与n-gram基线比较，在服务器托管日志数据上进行评估。

服务器托管日志数据和客户端设备拥有的缓存用于测量预测召回。尽管每一个都包含来自实际用户的数据片段，但客户机缓存被认为更准确地表示真实的类型数据分布。与日志不同的是，缓存数据的长度不会被截断，并且在Google拥有的应用程序中不受键盘使用的限制。因此，在Gboard的情况下，联邦学习能够使用高质量的训练数据。在表3中，计算了客户端缓存4的性能。引用的错误与用于联合计算的客户机数量直接相关。

|  |  |
| --- | --- |
| 模型 | 前1召回[%] |
| 纳克 | 12.5±0.2 |
| 服务器CIFG | 15.0±0.5 |
| 联合CIFG | 15.8±0.3 |

表4。在客户机拥有的数据缓存上评估服务器和联合CIFG模型与n-gram基线的预测召回。

模型性能也在实际生产实验中以Gboard用户子集为例进行了测试。与top-1召回相似，预测印象召回是通过将与用户输入的文本匹配的预测数量除以用户显示预测候选的次数来衡量的。预测印象回忆度量通常低于标准召回度量。零状态预测事件（用户打开Gboard应用程序但不提交任何文本）会增加显示次数，但不匹配。表5总结了在真实实验中印象回忆的表现。预测点击率（CTR）也在表6中给出，CTR定义为预测候选的点击数与建议的预测候选数的比率。所有结果的引用95%置信区间误差是使用带有用户桶的刀切法得出的。

|  |  |  |
| --- | --- | --- |
| 模型 | 前1召回[%] | 前三名召回率[%] |
| 纳克 | 5.24±0.02 | 11.05±0.03 |
| 服务器CIFG | 5.76±0.03 | 13.63±0.04 |
| 联合CIFG | 5.82±0.03 | 13.75±0.03 |

表5。在实时用户流量实验中评估了服务器和联邦CIFG模型与n-gram基线的预测印象召回。

|  |  |
| --- | --- |
| 模型 | 预测中心[%] |
| 纳克 | 2.13±0.03 |
| 服务器CIFG | 2.36±0.03 |
| 联合CIFG | 2.35±0.03 |

表6。服务器和联邦CIFG模型的预测CTR与n-gram基线的比较，在实时用户流量的实验中进行了评估。

对于服务器训练和联合训练，CIFG模型相对于基线n-gram FST模型改进了top-1和top-3召回。考虑到n-gram模型使用了数量级更大的词汇量，并且包括用户历史和联系人LMs等个性化组件，这些收益令人印象深刻。实时用户实验表明，CIFG模型还生成比n-gram预测高10%的点击率的预测。

结果还表明，联合CIFG在召回指标上的表现优于服务器训练的CIFG。表4显示，在对客户机缓存数据进行评估时，联合CIFG相对于服务器训练的CIFG将top-1召回提高了相对5%（绝对值为0.8%）。对服务器托管日志数据的比较表明，这两个模型的召回是可比较的，尽管日志并不能代表真正的类型分布。最重要的是，联合表的预测值比实际值提高了1%，而最重要的是，CIFG对顶层用户的预测值提高了1%。虽然这种比较并不完全是苹果对苹果-不同口味的SGD在每个训练环境中使用-结果表明，联合学习提供了一个更好的选择，以服务器为基础的神经语言模型的培训。

# 7.    结论

我们证明了一个使用联邦学习从头训练的CIFG语言模型在键盘下一个单词预测任务上的性能优于相同的服务器训练CIFG模型和基线n-gram模型。据我们所知，这是联邦语言建模在商业环境中的首批应用之一。联合学习通过在高度分布的计算设备中进行培训，同时提高语言模型的质量，为用户提供安全和隐私优势。

# 8.    致谢

作者要感谢googleai团队的同事们提供了联合学习框架并进行了许多有益的讨论。

# 9.    参考文献

[1] Monica Anderson，“技术设备所有权：2015年”http://www.pewinternet.org/2015/10/29/技术-设备-所有权-2015/，

访问日期：2018-10-02。

[2] Mehryar Mohri，“语言和语音处理中的有限状态传感器”，计算语言学，第23卷，第2期，第269-311页，1997年6月。

[3] Tom Ouyang、David Rybach、Franc¸oise Beaufays和Michael Riley，“使用有限状态传感器的移动键盘输入解码”，CoRR，第abs/1704.03987卷，2017年。

[4] Cyril Allauzen和Michael Riley，“移动语音输入的贝叶斯语言模型插值”，Interspeech 2011，2011，第1429-1432页。

[5] R.J.Williams和D.Zipser，“连续运行全递归神经网络的学习算法”，《神经计算》，第1卷，第2期，第270-280页，1989年6月。

[6] S.Hochreiter和J.Schmidhuber，“长短时记忆”，神经计算，第9卷，第8期，第1735-1780页，1997年11月。

[7] Yoshua Bengio，Rejean Ducharme，Pascal Vincent和“Christian Janvin”，神经概率语言模型，J.Mach。学习。研究，第3卷，1137-1155页，2003年3月。

[8] Yoon Kim、Yacine Jernite、David Sontag和Alexander M.Rush，“字符感知神经语言模型”，载于第三十届AAAI人工智能会议论文集。2016年，AAAI&apos;16，第2741-2749页，AAAI出版社。

[9] Ciprian Chelba、Tomas Mikolov、Mike Schuster、Qi Ge、Thorsten Brants、Phillipp Koehn和Tony Robinson，“用于衡量统计语言建模进展的10亿字基准”，发表于INTERSPEECH 2014，第15届国际言语传播协会年会，新加坡，2014年9月14日至18日，李海洲，Helen M.Meng，Bin Ma，Engsiong Chng，和Lei Xie，编辑，2014年，第2635-2639页，ISCA。

[10] Rafal Jozefowicz、Oriol Vinyals、Mike Schuster、Noam Shazeer和Yonghui Wu，“探索语言建模的极限”，2016年。

[11] Noam Shazeer，Azalia Mirhoseini，Krzysztof Maziarz，

Andy Davis、Quoc V.Le、Geoffrey E.Hinton和Jeff Dean，“惊人的大型神经网络：稀疏门控的专家层混合体”，CoRR，第abs/1701.065382017年。

[12] Brendan McMahan，Eider Moore，Daniel Ramage，Seth Hampson，and Blaise Aguera y Arcas，¨“从分散数据中高效学习深层网络”，载于第20届人工智能与统计国际会议论文集，2017年4月20日至22日，佛罗里达州劳德代尔堡，Aarti Singh和Xiaojin（Jerry）Zhu，编辑，2017年，机器学习研究论文集第54卷，第1273-1282页，PMLR。

[13] Ouais Alsharif，Tom Ouyang，Franoise Beaufays，Shumin Zik，Thomas Breuel，and Johan Schalkwyk，“键盘手势解码的长短时记忆神经网络”，2015年IEEE声学、语音和信号处理国际会议（ICASSP），第2076-2080页，2015年。

[14] Seunghak Yu，Nilesh Kulkarni，Haejun Lee，和Jihie Kim，“基于设备的神经语言模型的单词预测”，COLING 2018，第27届计算语言学国际会议：系统演示，新墨西哥州圣达菲，2018年8月，赵东燕编辑，2018年，第128-131页，计算语言学协会。

[15] Shaona Ghosh和Per Ola Kristensson，“键盘解码中文本校正和完成的神经网络”，CoRR，第abs/1709.0642922017年。

[16] Reza Shokri和Vitaly Shmatikov，“隐私保护深度学习”，载于第22届ACM SIGSAC计算机和通信安全会议记录，纽约，美国，2015年，CCS’15，第1310-1321页，ACM。

[17] Cynthia Dwork，“差异隐私”，第33届国际自动机学术讨论会，语言和编程，第二部分（ICALP 2006），意大利威尼斯，2006年7月，第4052卷，第1-12页，Springer Verlag。

[18] H.Brendan McMahan、Daniel Ramage、Kunal Talwar和Li Zhang，“在不损失准确性的情况下学习差异化的私人语言模型”，CoRR，第abs/1710.06963卷，2017年。

[19] 布伦丹·麦克马汉和丹尼尔·拉马奇，

“无需联合机器学习的集中式机器学习”https://ai.googleblog.com/2017/04/联合学习协作。html，访问日期：2018-10-04。

[20] 克劳斯·格雷夫（Klaus Greff）、鲁珀什·库马尔·斯里瓦斯塔瓦（Rupesh Kumar Srivastava）、扬·考顿（Jan Koutn´k）、巴斯特·施特布林克（Bas R.Steunebrink）和尤根·施密杜伯（Jurgen Schmidhuber），“LSTM:¨搜索空间奥德赛”，IEEE Trans。神经网络。学习系统，第28卷，第10期，2222-2232页，2017年。

[21]赵京贤，巴特·范梅里恩博尔，C¸阿格拉尔

Gulc¸ehre，Dzmitry Bahdanau，Fethi Bougares，Holger–Schwenk，Yoshua Bengio，“使用RNN编码解码器进行统计机器翻译的学习短语表示”，载于2014年自然语言处理经验方法会议论文集，EMNLP 2014，2014年10月25日至29日，卡塔尔多哈，SIGDAT会议，ACL的特殊利益集团，Alessandro Moschitti，Bo Pang和Walter Daelemans，Eds.2014，第1724–1734页，ACL。

[22]马丁阿巴迪、保罗·巴勒姆、陈建民、陈志峰、安迪·戴维斯、杰弗里·迪恩、马修·德文、桑杰·格马沃特、杰弗里·欧文、迈克尔·伊萨德、曼朱纳特·库德鲁尔、乔什·列文伯格、拉贾特·蒙加、雪莉·摩尔、德里克·戈登·默里、贝诺伊特·施泰纳、保罗·A·塔克、维杰·瓦苏德万、皮特·沃登、马丁·威克、袁宇、郑晓强，“Tensorflow:大规模机器学习系统”，第12届USENIX操作系统设计与实施研讨会，OSDI 2016，佐治亚州萨凡纳，美国，2016年11月2日至4日，Kimberly Keeton和Timothy Roscoe编辑，2016年，第265-283页，USENIX协会。

[23]Ofir Press和Lior Wolf，“使用输出嵌入改进语言模型”，载于计算语言学协会欧洲分会第15届会议论文集，EACL 2017，西班牙瓦伦西亚，2017年4月3日至7日，第2卷：短论文，Mirella Lapata，Phil Blunsom和Alexander Koller，2017年版，第157-163页，计算语言学协会。

[24]Hakan Inan、Khashayar Khosravi和Richard Socher，“捆绑词向量和词分类器：语言建模的损失框架”，CoRR，第abs/1611.01462卷，2016年。

[25]白宫，“网络世界中的消费者数据隐私：全球数字经济中保护隐私和促进创新的框架”，2013年1月。

[26]Keith Bonawitz、Vladimir Ivanov、Ben Kreuter、Antonio Marcedone、H.Brendan McMahan、Sarvar Patel、Daniel Ramage、Aaron Segal和Karn Seth，“用户持有数据上联邦学习的实用安全聚合”，在NIPS私人多方机器学习研讨会上，2016年。

[27]D.P.Kingma和J.Ba，“Adam:随机优化方法”，ArXiv电子印刷品，2014年12月。

[28]John Duchi，Elad Hazan和Yoram Singer，“在线学习和随机优化的自适应次梯度方法”，J.Mach。学习。研究，第12卷，第2121-2159页，2011年7月。

[29]Yurii Nesterov，“一种求解具有收敛速度（1/k2）的凸规划问题的方法”，Dokl。阿卡德。Nauk SSSR，第269卷，第543-5471983页。*o*

[[1]](" \l "_ftnref1" \o ")gboard.app.google.gl/获取

[[2]](" \l "_ftnref2" \o ")https://www.tensorflow.org/lite/