**分布式蒸馏技术在设备学习中的应用**

**摘要**

联合学习(FL)是一种机器学习设置，其中多个设备协作训练机器学习模型，同时保持训练数据的分散式存储。在当前的大多数训练方案中，中心模型是通过平均服务器模型的参数和来自客户端的更新参数来改进的。但是，只有当所有模型都具有相同的结构和大小时，才能直接平均模型参数，这在许多情况下可能是一个限制性约束。

在这项工作中，我们研究了更强大、更灵活的FL聚合方案。具体地说，我们提出了集成蒸馏用于模型融合，即通过客户未标记数据的模型的输出来训练中心分类器。该知识提炼技术在与基线FL算法在相同程度上

减轻了隐私风险和成本，但允许在例如大小、数值精度或结构上不同的异构客户端模型上进行灵活聚合。我们在各种CV/NLP数据集(CIFAR-10/100，ImageNet，AG News，Sst2)和设置(异构模型/数据)上进行了广泛的实证实验，结果表明，服务器模型的训练速度要快得多，所需的通信轮次比目前任何现有的FL技术都要少。

关键字：

培训 分散式 联合 代价 精馏

训练 分布式 联邦 损失 蒸馏

# 介绍

联合学习(FL)已经成为一种重要的机器学习范式，其中客户端的联合参与集中式模型的协作训练[63，51，66，8，5，42，34]。客户端将它们的模型参数发送到服务器，但从不发送它们的私有训练数据集，从而确保基本级别的私密性。联合培训的关键挑战之一是通信开销和延迟(人们希望用尽可能少的通信轮次来培训中心模型)，以及客户端异构性：培训数据(非I.I.D.-ness)以及硬件和计算资源在客户端之间可能会有很大变化，例如在商用移动设备上进行培训时。

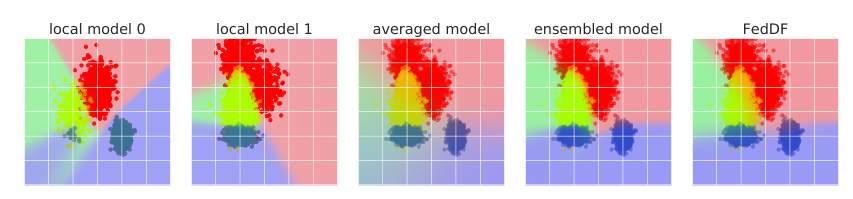
FL中的经典训练算法，例如联合平均(FEDAVG)[51]及其最近的自适应[53，44，25，35，26，58]，都基于参与客户端的参数的直接平均，因此仅当所有客户端的模型具有相同的大小和结构时才能应用。相反，集成学习方法[78，15，2，14，56，47，76]允许通过对单个模型的预测进行平均来组合多个异构弱分类器。然而，由于大量的参与客户端，直接在FL中应用集成学习技术在实践中是不可行的，因为它需要在服务器上保持所有接收到的模型的权重，并执行朴素的集成(Logits平均)来进行推理。

为了在更现实的环境中实现联合学习，我们建议使用集成蒸馏[7，22]进行鲁棒模型融合(FedDF)。我们的方案利用未标记的数据或人工生成的示例(例如，由GAN的生成器[17])来聚合来自所有接收到的(异构)客户端模型的知识。我们通过充分的实验结果证明，我们的集成蒸馏方法不仅解决了同构FL系统中存在的批处理归一化(BN)[31]的品质损失问题[24]，而且还可以打破异构客户模型之间的知识壁垒。我们的主要贡献是：

·提出了一种鲁棒的联邦模型融合的蒸馏框架，该框架允许异构的客户端模型和数据，并且对神经结构的选择具有鲁棒性。

·我们在各种CV/NLP数据集(CIFAR-10/100，ImageNet，AG News，Sst2)和设置(异构模型和/或数据)上进行了广泛的数值实验，结果表明，与任何现有的FL技术相比，服务器模型的训练速度要快得多，需要的通信轮数更少。

我们进一步提供了关于FedDF何时可以超越FEDAVG的见解(另见图1，它突出了基于参数平均的方法的内在局限性)以及哪些因素影响FedDF。



**图1** FEDAVG的局限性。我们考虑一个带有3层MLP的3类分类任务的有趣示例，并在输入空间上显示决策边界(RGB通道上的概率)。左边的两张图显示了单独训练的本地模型。右边的三个图评估聚合模型和全局数据分布；平均模型导致决策边界非常模糊。使用的数据集如图8(附录C.1)所示。

。

# 相关工作

**联合学习**。FL、FEDAVG[51]或LOCAL SGD[46]中的经典算法在所有设备都参与时，在多次本地SGD更新后对客户端模型执行加权参数平均，其权重与每个客户端的本地数据的大小成比例。基于客户流失的加权方案在[53，44]中进行了研究。为了解决直接平均模型参数的困难，[65，75]建议使用最优传输和其他对齐方案，在平均参数之前首先逐层对齐或匹配神经网络的单个神经元。然而，这些基于层的对齐方案要求客户端模型具有相同的层数和结构，这在实际的异构系统中是有限制的。另一项工作旨在改善本地客户培训，即由本地数据的异质性造成的客户漂移问题[43，35]。例如，FEDPROX[43]包含了本地培训的惩罚项。其他技术，如加速，最近出现在[25，26，58]中。

**知识蒸馏**。神经网络的知识蒸馏在[7，22]中首次引入。通过鼓励学生模型近似教师模型的输出逻辑，学生能够在边际质量损失的情况下模仿教师的行为[59，80，36，72，37，28，1，71]。一些工作研究了集合蒸馏，即将一组教师模型的知识提炼成一个学生模型。为此，现有的方法要么从教师模型集合[78，15，2，14]中取平均逻辑值，要么从特征层[56，47，76]中提取知识。这些方案中的大多数依赖于使用原始训练数据进行蒸馏过程。在真实数据不可用的情况下，最近的一些工作[54，52]表明，可以通过从教师模型的权重或通过与学生相反训练的生成器来制作伪数据来完成蒸馏。FedDF可以与所有这些方法结合使用。在这项工作中，我们考虑了用于集成蒸馏的未标记数据集，这些数据集可以从其他领域收集，也可以直接从预先训练的生成器生成。

**与最近的联邦学习工作进行比较**。Guha等人。[18]针对支持向量机的损失目标，提出了通过未标记数据进行“一次融合”的方法，同时考虑了不同神经结构和任务的多轮场景。FD[33]利用蒸馏来降低FL通信成本。为此，FD同步在本地训练期间累积的每个标签。然后，每个标签的平均对数(在本地步骤和客户端上)将被用作下一轮本地培训的蒸馏正规化。与FEDAVG相比，FD在MNIST上的品质下降了大约15%。相比之下，FedDF显示出比FEDAVG更优越的学习性能，并且可以显著减少通信轮次，从而在各种具有挑战性的任务上达到目标精度。FedMD[41]和最近提出的Cronus[9]考虑通过公共数据集上每个样本的平均对数来学习。在对已标记的公共数据集进行初始预训练之后，FedMD迭代地学习公共和私有数据集以进行个性化，而在Cronus中，公共数据集(带有软标签)与本地私有数据联合用于本地训练。由于FedMD同时在标记为公共和私有的数据集上训练客户端模型，因此模型分类器必须包括这两个数据集中的所有类。Cronus在其协作培训阶段混合了公共和私人数据，用于当地培训。因此，对于这些方法，公共数据集的构建需要仔细考虑，甚至需要事先了解客户的私有数据。此外，这些修改对当地培训质量的影响尚不清楚。FedDF没有面临这样的问题：我们证明了FedDF对蒸馏数据集的选择是鲁棒的，并且蒸馏是在服务器端执行的，而本地训练不受影响。我们在附录A中包括了与FedMD，Cronus的详细讨论。在准备这个版本时，我们还注意到了其他当代作品[69，10，82，19]，我们将讨论推迟到附录A。。

# 用于鲁棒模型融合中的集成蒸馏技术

现。

## 3.1 技术预展示

我

## 3.2 最优解的特性

我

。

# L2GD: Loopless Local GD

在

。

## 局地GD动态变化和平均步长

此

。

# 4.3收敛理论

我

。

# 4.4优化速率和通信

让

。

# Loopless Local SGD with Variance Reduction

正

。

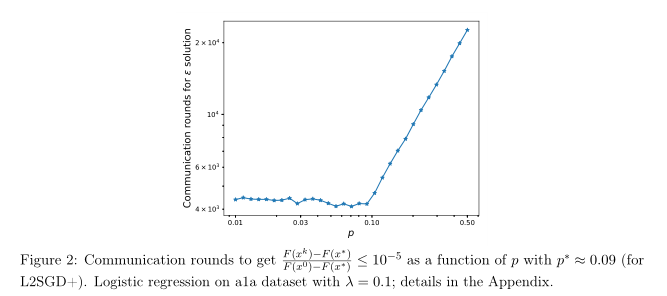
## 设置

。

## Theory

现

。



**图2**

# 实验

在

。

# 结论

我。

# 参考文献

[1] J. Koneˇ cn` y, H. B. McMahan, F. X. Y u, P . Richtárik, A. T. Suresh, and D. Bacon, “Federated learning: Strategies for improving communication efficiency,” arXiv preprint arXiv:1610.05492,2016.