读DENSE: Data-Free One-Shot Federated Learning

这篇论文关注一次性联邦学习 (One-Shot Federated Learning, OFL)，着重解决现有OFL方法中的数据依赖和模型同构性假设问题，提出了一种无需数据 (Data-Free) 的方法。

现有的一次性联邦学习方法大多不实用或面临固有限制，例如，需要公共数据集、客户端模型必须是同构的、需要上传额外的数据/模型信息。为了克服这些问题，本文提出了一个新颖的联邦学习框架DENSE，它通过一个数据生成阶段和一个模型蒸馏阶段来训练全局模型。
DENSE由于以下优点可以在现实中应用：

1. 与其他方法相比，DENSE不需要在客户端和服务器之间传输额外信息（除了模型参数）；
2. DENSE不需要任何辅助数据集进行训练；
3. DENSE考虑了FL中的模型异构性，即不同的客户端可以拥有不同的模型架构。

引言

这里首先介绍了一次性联邦学习的优点和有待提高之处。

一次性联邦学习是一个有前景的解决方案，有这样几个好处：

1. 多轮训练在某些场景下不实用，例如模型市场，用户只能从市场购买预训练模型而没有任何真实数据。
2. 频繁的通信带来了很高的被攻击风险。例如，频繁通信很容易被攻击者拦截，他们可以发起中间人攻击，甚至从梯度中重构训练数据。这样，一次性FL由于其单轮特性，可以降低被恶意攻击者拦截的概率。

然而，现有的一次性联邦学习方法均存在局限性：

- 数据集蒸馏会产生额外的通信成本和潜在的隐私泄露风险。
- 基于聚类的方法需要将聚类中心上传到服务器，从而产生额外的通信成本。
- 这些方法都没有考虑模型异构性，即不同客户端拥有不同的模型架构。

对此，本文提出了DENSE框架，它通过一个数据生成阶段和一个模型蒸馏阶段来训练全局模型：在第一阶段，利用集成模型（即客户端上传的本地模型的集成）来训练一个生成器，该生成器可以生成用于第二阶段训练的合成数据。在第二阶段，将集成模型的知识蒸馏到全局模型中。与基于FedAvg的传统FL方法相比，该方法不需要对模型参数进行平均，因此它可以支持异构模型，即客户端可以拥有不同的模型架构。

方法：无需数据的一次性联邦学习 (Data-Free One-Shot Federated Learning)

数据生成阶段

第一阶段目标是训练一个生成器来生成合成数据。生成器应该能生成与客户端训练数据具有相似分布的数据。现有的研究通过利用预训练的GAN生成数据，然而，预训练的GAN是在公共数据集上训练的，其数据分布可能与客户端的训练数据不同。此外，需要考虑模型异构性，这使得问题更加复杂。
为了解决这些问题，本文将训练一个同时考虑相似性、稳定性、和可迁移性的生成器。具体来说，给定一个随机噪声z（从标准高斯分布生成）和一个随机标签y（从均匀分布生成），生成器旨在生成合成数据x = G(z)，使得x与客户端的（带有标签y的）训练数据相似。个人感觉这里和difussion有些类似。
首先，考虑合成数据x和训练数据之间的相似性。由于无法访问客户端的训练数据，我们不能直接计算合成数据和训练

数据之间的相似性。相反，我们首先计算x由集成模型计算出的平均logits（即最后一个全连接层的输出）。
$D(x;\theta^k k=1^m)=(1/m)\Sigma k\in Cf^k(x;\theta^k)(1)$
其中 m = |C|（客户端数量），$D(x;\theta^k)$ 是 x 的平均logits，θ^k 是第k个客户端的参数。而 $f^k(\cdot;\theta^k)$ 是客户端k的预测函数，输出给定参数 θ^k 的logits。为简单起见，在本文的其余部分使用 D(x) 表示 $D(x;\theta^k)$。
然后，用以下交叉熵（CE）损失来最小化平均logits和随机标签y:
$L_CE(x,y;\theta_G)=CE(D(x),y)$
在训练阶段，D(x) 和 y 之间的损失可以轻易降低到几乎为0，这表明合成数据与集成模型完美匹配。
然而，仅利用CE损失是无法获得高性能的，这可能因为集成模型是在非独立同分布(non-IID)数据上训练的，生成器可能不稳定并陷入局部最有或对合成数据过拟合。

其次，为了提高生成器的稳定性，本文建议添加一个额外的正则化来稳定训练。
$L_BN(x;\theta_G)=(1/m)\Sigma_{k\in C}\Sigma_l(||\mu_l(x)-\mu_{k,l}||^2+||\sigma_l^2(x)-\sigma_{k,l^2}||)(3)$
其中 $\mu_l(x)$ 和 $\sigma_l^2(x)$ 是对应于生成器$G(\cdot)$的第l个BN层的批处理均值和方差估计，$\mu_{k,l}$ 和 σ_{k,l^2} 是第k个客户端模型 $f^k(\cdot)$ 的第l个BN层的均值和方差。BN损失最小化了合成数据的特征图统计量与客户端训练数据的特征图统计量之间的距离。因此，无论数据是非独立同分布还是独立同分布，合成数据都可以具有与客户端训练数据相似的分布。

通过利用CE损失和BN损失，我们可以训练一个能生成合成数据的生成器，但我们观察到合成数据可能远离集成模型的决策边界，这使得集成模型难以将其知识转移给全局模型。
为了解决这个问题，本文主张生成更多落在集成模型和全局模型决策边界之间的合成数据。位于决策边界同一侧的合成数据，对学习全局模型帮助较小。位于决策边界之间的合成数据，全局模型和集成模型对这些数据有不同的预测，可以帮助全局模型更好地学习集成模型的决策边界。

受上述观察的启发，本文引入了一个新的边界支持损失 (boundary support loss)，它促使生成器生成更多位于集成模型和全局模型决策边界之间的合成数据。
具体而言，将合成数据分为两组：

1. 全局模型和集成模型对第一组数据有相同的预测
   $(argmax_cD^c(x)=argmax_cf_s^c(x;\theta_s))$
2. 对第二组数据有不同的预测
   $(argmax_cD^c(x)\neq argmax_cf_s^c(x;\theta_s))$
   其中 $D^c(x)$ 和 $f_s^c(x;\theta_s)$ 分别是集成模型和全局模型对应第c个标签的logits。第一组数据位于两个决策边界的同一侧，而第二组数据位于决策边界之间。这里使用Kullback-Leibler（KL）散度损失来最大化全局模型和集成模型在第二组数据上预测的差异：
   $L_div(x;\theta_G)=-\omega KL(D(x),f_s(x;\theta_s))$
   其中 $KL(\cdot,\cdot)$ 表示KL散度损失，$\omega=I(argmax_cD^(c)(x)\neq argmax_cf_s^c(x;\theta_s))$ 对于第一组数据输出0，对于第二组数据输出1，I(a)是指示函数，如果a为真则输出1，否则输出0。通过最大化KL散度损失，生成器可以生成更多对模型蒸馏阶段更有帮助的合成数据，并进一步提高集成模型的可迁移性。
   通过结合上述损失，我们可以得到生成器损失如下：
   $L_gen(x,y;\theta_G)=L_CE(x,y;\theta_G)+\lambda_1L_BN(x;\theta_G)+\lambda_2L_div(x;\theta_G)$
   其中 λ₁ 和 λ₂ 是缩放因子。

模型蒸馏阶段

在第二阶段，我们使用（前一节讨论的）生成器和集成模型来训练全局模型。先前的研究表明，模型集成提供了一种提高学习模型准确性和稳定性的通用方法。一个直接的方法是通过聚合所有客户端模型的参数来获得全局模型（例如，通过FedAvg）。然而，在实际应用中，客户端很可能拥有不同的模型架构，使得FedAvg无效。此外，由于不同客户端的数据是非独立同分布的，FedAvg无法提供良好的性能甚至可能发散。
为此，这里将集成模型的知识蒸馏到全局模型中，通过在相同的合成数据上最小化集成模型（教师）和全局模型（学生）之间的预测差异。
首先，我们根据公式(1)计算合成数据的平均logits，即 $D(x)=(1/m)\Sigma_{k\in C}f^k(x;\theta^k)$。与无法聚合异构模型的传

统聚合方法（如FedAvg）相比，平均logits可以轻松应用于异构和同构FL系统。
然后，我们使用平均logits通过最小化以下目标函数来蒸馏集成模型的知识：
$L_dis(x;\theta_s) = KL(D(x), f_s(x;\theta_s))(6)$
通过最小化KL损失，我们可以在不考虑数据和模型异构性的情况下，用集成模型的知识和合成数据训练一个全局模型。
也就是说，DENSE对客户端的本地模型没有限制，即客户端可以使用任意技术训练模型。因此，DENSE是一种兼容的方法，可以与任何本地训练技术相结合，以进一步提高全局模型的性能。

关于隐私保护的讨论

研究表明，恶意用户可能使用GANs来重构其他参与者私有数据集的样本，从而发起攻击。此外，在FL中，服务器和客户端之间交换模型可能导致潜在的隐私泄露。DENSE禁止生成器直接看到真实数据，并且只有一轮通信，这降低了隐私泄露的风险。

关于FL中知识蒸馏的讨论

在传统的FL框架中，所有用户必须就全局模型的特定架构达成一致。曾有人通过通过使用代理数据集进行知识蒸馏，这的确缓解了由non-IID数据引起的模型漂移问题。然而，对代理数据的要求使得这种方法在许多应用中不切实际，因为精心设计的数据集并不总是可用。
无需数据的知识蒸馏是一种有前景的方法，可以在没有任何真实数据的情况下将教师模型的知识转移给学生模型。也有研究通过在每轮通信中生成合成数据进行模型融合的无需数据集成蒸馏，但这需要高昂的通信成本和计算成本。在本文中，则更关心在异构模型的情况下，通过仅一轮通信获得一个好的全局模型，这更具挑战性和实用性。
也有研究通过将从本地模型的预测派生出来的生成器广播给所有客户端，然后客户端需要将他们的生成器发送到服务器，但是这增加了通信负担。更严重的是，生成器可以直接访问本地数据（生成器可以轻易记住训练样本），这可能引起隐私担忧。由于本文中使用的生成器始终存储在中央服务器，它永远不会看到任何真实的本地数据。

实验

结果

在真实世界数据集上的评估

为了评估我们方法的有效性，我们在不同的non-IID设置下（通过改变α={0.1, 0.3, 0.5}）进行了实验，并在表1中报告了不同数据集和不同方法的性能。结果显示：

1. DENSE在所有数据集上都达到了最高的准确率。特别是，在CIFAR10数据集上当α=0.3时，DENSE比最佳基线方法Fed-ADI高出5.08%。
2. FedAvg的性能最差，这意味着在non-IID设置下直接平均模型参数无法在一次性FL中获得良好性能。
3. 随着α变小（即数据变得更不平衡），所有方法的性能都显著下降，但是即使在高度倾斜的设置下，DENSE仍然显著优于其他方法。

模型蒸馏的影响

当E=400时，DENSE优于每个客户端的本地模型，而FedAvg的表现不如每个客户端的本地模型。这验证了模型蒸馏可以增强训练，而直接聚合在non-IID设置的一次性FL训练中是有害的。

在异构FL中的结果

本实验在CIFAR10数据集上应用了五种不同的CNN模型，使用Dirichlet分布 α = {0.1, 0.3, 0.5}。
异构模型包括：

1. 一个ResNet-18
2. 两个小型CNN：CNN1和CNN2
3. 两个Wide-ResNets (WRN)：WRN-16-1和WRN-40-1
   对于知识蒸馏，这里使用ResNet-18作为服务器的全局模型。结果表明，在non-IID数据分布和不同模型架构设置下的FL是一个相当具有挑战性的任务。不过即使在这种设置下，DENSE仍然有显著优势。

方法分析

客户端数量的影响

随着客户端数量m的增加，所有方法的准确率都会下降，但是DENSE仍然有显著优势。

与不平衡学习的结合

将我们的方法与这些解决不平衡本地数据的技术相结合可以带来更有效的FL系统，例如LDAM。本实验在CIFAR10和SVHN数据集上比较了原始DENSE和与LDAM结合的DENSE（DENSE+LDAM）在 α = {0.1, 0.3, 0.5} 下的性能，结果是通过在高度倾斜的数据上将DENSE与LDAM结合可以实现显著的改进。

合成数据的可视化

这里的可视化表明合成数据与原始数据不相似，这可以有效降低泄露客户敏感信息的可能性。尽管合成数据看起来与原始数据大不相同，DENSE通过使用这些合成数据进行训练仍然比其他基线方法获得了更高的性能。

扩展到多轮

随着通信轮数 T_c 的增加，DENSE的性能提高，并且在 T_c = 5 时达到最佳性能。这表明DENSE可以扩展到多轮FL，并且可以通过增加通信轮数来进一步提高性能。

L_BN 和 L_div 的贡献

本实验研究了用于数据生成的不同损失函数的贡献。这里进行了"留一法"测试，并通过移除 L_div 和移除 L_BN 来展示结果。此外，也进行了移除 L_div 和 L_BN，即仅使用 L_CE 的结果。结果表明，仅使用 L_CE 训练生成器导致性能不佳。此外，移除 L_BN 损失或 L_div 损失也会影响全局模型的准确性。这些损失函数的组合导致了全局模型的高性能，这表明损失函数的每个部分在增强生成器方面都起着重要作用。