通过联邦学习中的非真实蒸馏来保存全局知识

Gihun Lee*、**Minchan Jeong***、**Yongjin Shin**、**Sangmin Bae**、**Se-Young Yun**
韩国科学技术院
{opcrisis, mcjeong, yj.shin, bsmn0223, yunseyoung}@kaist.ac.kr

摘要

在联合学习中，通过聚合客户在本地训练的模型来协作学习强大的全局模型。尽管这排除了直接访问客户数据的需要，但全局模型的融合通常会受到数据异构性的影响。这项研究从类比到持续学习开始，并表明*遗忘*可能是联邦学习的瓶颈。我们观察到全局模型忘记了前几轮的知识，而局部训练导致忘记了局部分布之外的知识。根据我们的发现，我们假设解决遗忘问题将缓解数据异质性问题。为此，我们提出了一种新颖有效的算法， *Federated Not-True Distillation(FedNTD)，它只为非真实*类别保留对本地可用数据的全局视角。在实验中，FedNTD 在各种设置上展示了最先进的性能，而不会损害数据隐私或产生额外的通信成本 s [I] [II] [II]。[III] [IV]

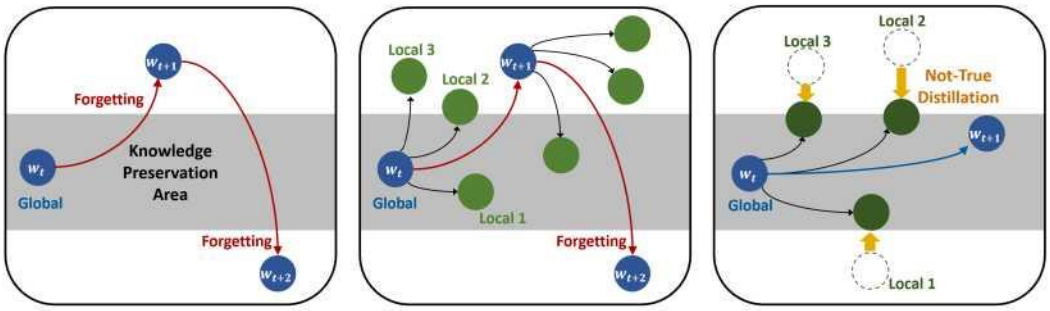
1 简介

目前，手机、车辆、设施等边缘设备正在收集海量数据。由于数据可能分布在众多设备上，因此通常需要分散训练来训练深度网络模型。联合学习[23 , 24]是一种分布式学习范例，可以在保护客户数据隐私的同时学习全局模型。在联邦学习中，客户端使用他们的私有数据独立训练本地模型，然后服务器将它们聚合成一个全局模型。在此过程中，大部分计算由客户端设备执行，而全局服务器仅聚合模型参数并将其分发给客户端[2 , 50] .

大多数联邦学习算法都基于 FedAvg [37]，它通过与每个客户端拥有的本地数据量成比例的加权平均来聚合本地训练的模型参数。虽然到目前为止已经提出了各种联邦学习算法，但它们各自以某种方式进行参数平均[1 , 9 , 20 , 30 , 47 , 52]。尽管这种聚合方案在经验上运作良好，并在所有客户端设备都处于活动状态且 iid 分布式（又名 LocalSGD）时提供了概念上理想的框架，但数据异构性问题[31] ， 56]是联邦学习应用程序的一个臭名昭着的挑战，并阻止了它们的广泛适用性[19 , 29]。

由于客户端生成自己的数据，因此数据分布不均。更准确地说，跨客户端的本地数据是从异构的底层分布中提取的；因此，本地可用数据无法代表整体全球分布，这被称为数据异质性。尽管在许多现实世界场景中不可避免地会发生数据异构性，但它不仅使理论分析变得困难[31 , 56] ，而且还会降低许多联邦学习算法的性能[18 , 27]。通过解决数据异质性问题，学习对部分参与变得更加稳健[31 , 37]，并且通信成本也因更快的收敛而降低[46 , 52]。

* 平等贡献。
[II] https://github.com/Lee-Gihun/FedNTD 第 36 届神经信息处理系统会议 (NeurIPS 2022)。



(a) Continual Learning (b) Federated Learning (FedAvg) (c) Federated Learning (FedNTD)

图 1：学习场景中的遗忘概述。作为 (a) 持续学习中的灾难性遗忘，(b) 联邦学习也会经历遗忘。然而，(c) FedNTD 通过在局部训练期间保留全局知识来防止遗忘。

有趣的是，持续学习[42 , 45]面临着类似的挑战。在持续学习中，学习者模型在一系列任务上不断更新，目标是在整个任务上表现良好。不幸的是，由于每个任务的异构数据分布，对任务序列的学习通常会导致*灾难性的遗忘*[36、40] ，从而适应新任务会干扰对先前任务很重要的参数。因此，模型参数会偏离希望保留先前知识的区域（图 1a ）。

我们的第一个猜想是联邦学习中也存在这种遗忘。虽然服务器聚合了本地模型，但它们训练的分布可能与前几轮有很大不同。因此，全局模型在每一轮都面临分布变化，这可能会导致持续学习中的遗忘（图 1b ）。为了凭经验验证这个类比，我们检查了全局模型的预测一致性。更具体地说，我们在通信轮次进行时测量其类别准确性。

观察结果验证了我们的猜想：全局模型的预测在通信轮次之间高度不一致，显着降低了先前模型原本预测良好的某些类别的预测性能。我们更深入地分析了对局部更新的参数进行平均是如何导致这种遗忘的，并确认它发生在局部训练中：对应于局部分布之外的区域的全局知识很容易被遗忘。由于仅对本地模型进行平均无法恢复它，因此全局模型难以保留先前的知识。

根据我们的发现，我们假设减轻遗忘问题可以缓解数据异质性（图 1c ）。为此，我们提出了一种新算法 Federated Not-True Distillation (FedNTD)。FedNTD 利用全局模型对本地可用数据的预测，但仅针对不真实的类别。我们展示了 FedNTD 对在本地分布之外保存全局知识的影响及其对联邦学习的好处。实验结果表明，FedNTD 在各种设置中实现了最先进的性能。

总而言之，我们的贡献如下：

- 我们对联邦学习中的遗忘进行了系统研究。局部分布之外的全局知识很容易被遗忘，并且与数据异构性问题密切相关（**第 2 节**）。
- 我们提出了一种简单而有效的算法 FedNTD 来防止遗忘。与之前的作品不同，FedNTD 既不会损害数据隐私，也不会产生额外的通信负担。我们验证了 FedNTD 在各种设置上的功效，并表明它始终如一地实现了最先进的性能（**第 3 节，第 4 节**）。
- 我们分析了 FedNTD 如何使联合学习受益。FedNTD 的知识保存改善了局部训练后的权重对齐和权重发散（**第 5 节**）。

1.1 预赛

联邦学习我们的目标是在由 K *个客户端和中央服务器组成的联邦学习系统中训练图像分类模型。每个客户端 k 都有一个本地数据集 D k，其中整个数据集* **D** = |J ₖ Pr ĸ s **D** ᵏ。在每个通信轮 *t*，服务器分发当前全局 22

模型参数 $w^{(t_q}$到采样的本地客户端 $K^{(t)}$。从 $w^{(t_q}$开始，每个客户端 k e $K^{(t)}$使用其本地数据集 D^k更新模型参数 w_{ptt}，目标如下：

$$w^{(t} = \text{argmin}\ E_{xy,.\ D}[\ \underset{w}{\mathbf{L}}(w; w^{(t-1)}, x, y)] \ . \tag{1}$$

其中 L 是损失函数。*在第 t* 轮结束时，采样客户端将本地更新的参数上传回服务器，并通过参数平均聚合为 $w^{(t)}$，如下所示：

$$w^{(t)} = \sum_{k=1}^{K} \frac{|D_k|}{|D|} w_k^{(t)} \tag{2}$$

知识蒸馏给定教师模型 *T* 和学生模型 *S*，知识蒸馏[17]使用温度 *r* 匹配它们的软化概率 q_T 和 q_S。q_r的第 c 个值可以描述为 $q_r(c) = \frac{\exp(z_c/r)}{\sum_i \exp(z_i/r)}$，其中 z_c是 logits 向量 *z* 的第 c 个值，*C* 是数字类。给定样本 *x*，学生模型 S 是通过 使用超参数 ^的交叉熵损失 $\mathbf{L}_{CE}$ 和 Kullback - Leibler 散度损失 L_{KL} 的线性组合来学习的：

$$L = (1 - £)\, L_{CE}(q, 1_y) + £ r^2 L_{KL}(q_T, q_T) \tag{3}$$

$$L_{CE}(q; 1_y) = -\sum_{c=1}^{C} 1_y(c) \log q(c); \qquad L_{KL}(q_{TS}, qT) = -\sum_{c=1}^{C} q_T(c) \log \frac{q'T(c)}{q_T(c)} \tag{4}$$

2 联邦学习中的遗忘

为了解非独立同分布数据如何影响联邦学习，我们对异质局部进行了一项实验研究。我们选择 CIFAR-10 [25]和[37]中的四层卷积神经网络。我们使用 Latent Dirichlet Allocation (LDA) 将数据拆分为 100 个客户端，通过 $p \sim Dir(a)$ 将 c 类样本的分区分配给客户端。异质性水平随着 *a 的* 降低而增加。我们使用 FedAvg 训练模型进行 200 轮通信，并在每轮针对 5 个本地时期优化 10 个随机抽样的客户端。更多详细信息在附录 B 中。

2.1 全局模型预测一致性

为了证实我们关于遗忘的猜想，我们首先考虑全局模型的预测如何随着通信轮次的进行而变化。如果数据异质性导致遗忘，则更新后的预测（即参数平均）与上一轮相比可能不太一致。为了检查它，我们观察模型在每一轮的类测试精度，并测量它与前一轮的相似性。结果在图 2a 和图 2b 中提供。

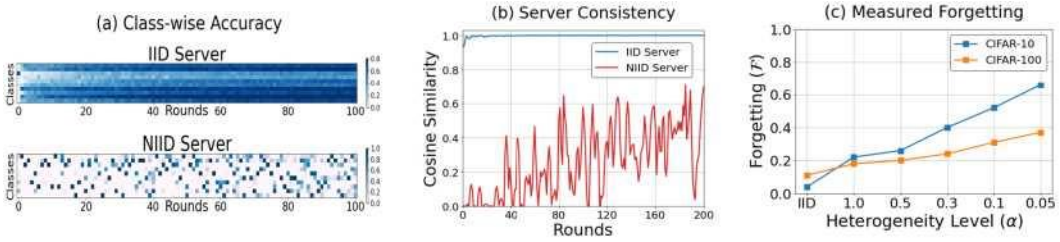


图 2：对全局服务器模型的遗忘分析。(a)：CIFAR-10 IID 和 NIID (a=0.1) 案例的类别测试准确度。(b)：在 IID 和 NIID (a = 0.1) 情况下，类准确度向量与前一轮全局模型的余弦相似度。(c)：在 CIFAR-10 和 CIFAR-100 上通过不同的异质性水平遗忘 *F*。

正如预期的那样，虽然从 iid locals（IID 服务器）学习的服务器模型在每一轮中均匀地预测每个类，但在非 iid 情况下（NIID 服务器）的预测是高度不一致的。在非独立同分布的情况下，先前全局模型原本预测良好的某些类的测试精度往往会显着下降。这意味着遗忘发生在联邦学习中。

为了衡量遗忘与数据异质性的关系，我们借用了*向后迁移*(BwT) [5]的概念，这是持续学习中普遍存在的遗忘度量[4、7、8、14]，如下所示：

$$F = cs \sum_{c=1}^{\text{摄氏度}} \max_{te\{1,\ldots T-if^c} (\mathbf{A}_{,(<)} - \mathbf{A}^{T}q)_c \tag{5}$$

其中 $_{Act}$ 是第 t 轮 c 类的准确率。请注意，遗忘度量 F 捕获了学习结束时每个类的峰值准确度与最终准确度之间的平均差距。图 2c 绘制了不同异质性水平的结果，表明随着异质性水平的增加，全局模型的遗忘更加严重。

2.2 本地分布之外的知识

我们仔细研究本地培训，以调查为什么聚合本地模型会导致遗忘。在持续学习的观点中，一种直接的方法是观察新分布的拟合如何降低旧分布的性能。然而，在我们的问题设置中，本地客户可以有任何类别。鉴于它们在本地分布中的部分在客户端之间有所不同，因此这种严格的比较是棘手的。因此，我们制定*了局部分布* p(**D**) 及其*局部分布 p)(D')*来系统地分析局部训练中的遗忘。

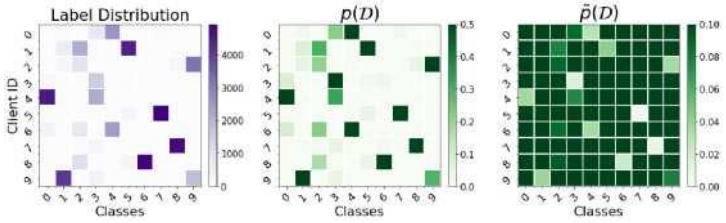


图 3：CIFAR-10 (a = 0.1)上的局部分布 p(D) 和局部分布 *p(D)示例。*

定义 1.考虑*在 C 类分类问题中由 N 个数据点 x $_i$ 及其标签 y $_i$ 组成的局部数据集 D。内部**分布向量** p ^{k} = p(D ^{k})及其**外部分布向量** p ^{k} = p(**D** ^{k})是*

$$p = [\,_{pi\,i} \quad ,:::,P_C\,], \quad \text{其中 } p_c := Ah^* = c) \overset{1}{\text{牛顿}} \quad (6)$$

$$p = [\mathrm{p}_{\,i} \quad ,...pp_c\,], \quad \text{其中 } p_{\,c} \quad := \mathrm{C}\,'\,|\,(1 - \quad \mathrm{P}_c) \qquad (7)$$

外局部分布 *p(D)*的基本思想是将较高比例分配给局部数据集中样本较少的类。因此，它对应于局部分布 p(**D**) 无法表示的全局分布中的区域。请注意，如果 p(**D**) 是均匀的，则 p(**D**) 也会折叠成均匀的，这在直觉上对齐得很好。图 3 提供了 10 个客户端及其 p(**D**) 和 p(**D**) **的标签分布示例**。

如图 4 所示，我们在每一轮通信中测量全局和局部模型在 p(**D**) 和 p(**D**) **上的准确度变化**。经过局部训练后，局部模型非常适合 p(**D**)（图 4a），聚合的全局模型在其上也表现良好。另一方面，p(**D**) 上的精度显着下降，其上的全局模型精度也下降（图 4b）。

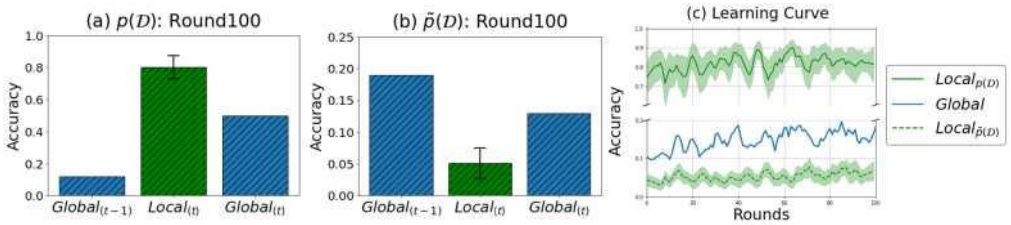


图 4：CIFAR-10 (a=0.1) 上 p(D) 和 p(D) 的全局模型和采样局部模型的准确性。误差条代表 10 个抽样本地客户的标准偏差。在 (a) 和 (b) 中，p(D) 和 p(D) 的全局模型精度是根据 10 个样本客户的联合分布测量的。

总而言之，关于局部外分布 p(**D**) 的知识很容易在局部训练中被遗忘，从而导致全局模型的遗忘。根据我们的发现，我们假设遗忘可能是联邦学习的绊脚石。

2.3 遗忘和局部漂移

首先由[56]根据经验观察到，局部更新偏离理想的全局方向已被广泛讨论为异构联邦学习中收敛缓慢且不稳定的主要原因[21,30,31]。不幸的是，考虑到分析这种漂移的困难，一种常见的方法是假设局部函数梯度之间存在有界差异[20、31]。

局部外分布知识保存的一个有趣特性是它可以将局部梯度修正为全局方向。我们定义梯度多样性 A 来衡量局部梯度的相异性，并说明知识保存的效果如下：

定义 2.*对于均匀加权的 K 个客户端，局部函数 f^k 对全局函数 f = -1 2 K _ 1 f ᵏ的梯度多样性 A 定义为:*

$$\frac{\frac{K}{K}\ \mathbf{i}\ \mathbf{IV}^{k}\ \mathbf{II}^{2}}{\text{四 二}^{2}} \tag{8}$$

这里，A ￥ 1 测量局部函数 fk s 与全局函数 f 的梯度方向对齐。请注意，随着局部函数梯度 Vfks 的方向变得相似，A 变得更小——**例如**，如果 ||Vfk ‖的大小 ² s 固定，**V** f ᵏ s 的方向相同时，A 最小。为了理解保留知识对局部外分布 p(**D**) **的影响，我们通过在 p(D)** 上添加具有因子 *ft* 的梯度信号来分析局部梯度及其多样性如何变化，并得到以下命题。

命题 1.假设具有本地分布 p ^{k} = *[p* ᵏ *,... ,p* C *]的均匀加权 K 个客户端。如果我们假设类梯度 g* ₀*是正交的且幅度均匀，则增加 ft* =C *C* /2 — 1 *会降低局部梯度 V f* ^{k} = *(p* ^{k} + ftp ᵏ) • g 的梯度多样性 A，*比例为:*

$$\frac{BA}{Bft} \qquad \frac{MK,C,P}{(1+\text{ft})^{2}.} \tag{9}$$

其中 ft 代表知识保存对局部外分布 p ^{k} *的影响。M* ₖ*, c* ,ₚ > 0 *是常数项，由 K、C 和 (p* ^{k} *ft* ₋₁ *,*

证明在附录 P 中给出。请注意，这里我们将 f ᵏ视为类损失的总和 ₀ p ₖ **L** ₀ = E ₓ | ᵧ _ ₀*[L(x;* w)] 是特定类别 c 上的损失。当 ft = 0 时，没有正则化来保留外分布知识，因此局部模型只需要拟合内局部分布 p^k。上述命题表明，关于局部分布 p ᵏ的保留知识 （即，随着 ft 的增加）引导局部梯度方向与全局梯度更加对齐，从而降低梯度多样性 A。这种遗忘视角提供了在模型预测级别处理数据异质性的机会。

3 FedNTD：联合非真实蒸馏

在本节中，我们提出联合非真实蒸馏 (FedNTD) 及其主要特征。FedNTD 的核心思想是只为非真实类保留全局视图。更具体地说，FedNTD 通过交叉熵 l o ss **L** CE 和非真实蒸馏损失 **L** NTD 之间的线性组合损失函数**L** 进行局部蒸馏：

$$\text{ᴸ} = \mathbf{L}_{\text{CE}}(q^{l};1_{\text{y}}) + \text{ft} \bullet \mathbf{L}_{\text{NTD}}(q^{l}{}_{\text{T}};\sim g)_{\text{。}} \tag{10}$$

这里，超参数 ft 代表知识在局部分布上的保存强度。然后，非真实蒸馏损失 **L** NTD 定义为非真实 softmax 预测向量 $q^{l}{}_{\text{T}}$和 q_{g}之间的 KL-Divergence 损失，如下所示：

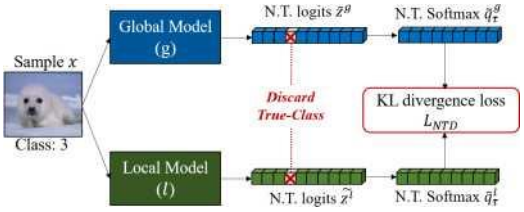


图 5：Not-True Distillation 概述。在 softmax 中忽略了真实类（*第 3 类*）logits。

$$L_{\text{NTD}}(\sim_{\text{T}},q^{g}) = -\wedge\sim_{\text{T}}(c)\log_{c=1,\underline{c\wedge\mathbf{y}}}^{C}\frac{qT(c)}{q_{\text{T}}(c)} \quad \text{where} \quad qT(c) = \frac{\exp(z_{g}/\tau)\underline{\text{SC}\wedge\mathbf{y}}\ \text{exp}}{(z\sim/\tau_{q})} \quad qT(c) = \frac{\exp(z_{g}/\tau)\underline{\text{SCft}\mathbf{y}}\ \text{exp}}{(z_{c^{g}}/\tau_{q})} \quad (@c\ /\ y). \tag{11}$$

仅对不真实的类 logits 采用带有温度 T *的 SOFTMAX*。图 5 *说明了在给定样本* x 的情况下，不真实的蒸馏是如何工作的。注意忽略真实类 logits 使得 $\mathbf{L}_{NTD}$ 到真实类的梯度信号为 0。详细算法在算法 1 中提供。

算法 1 联合非真实蒸馏 (FedNTD)

输入： 总轮数 T，局部时期 E，数据集 $\mathbf{D}$，样本客户集 $K_{ptq} \in K$ 在第 t 轮， 学习率 7

为全局服务器权重初始化 w_{p} 0)

对于每一轮通信 t = 1, • • • , T **do**

  服务器采样客户端 K p t q 并广播 W p t q ^ $w_{pt'q}$

  对于每个客户端 k e K $_{p}{}^{tq}$ **并行做**

    对于局部步骤 e = 1 • • • E **do**

      对于批次 j = 1 • • • B **do**

        w ktq ^ w ktq — 7r »L(w ktq ; [D k] j)   使用[方程 10]

      结束于

    结束于

  结束于

  上传 *wk 到*服务器

  服务器聚合' w (t+1) ^ i Kh j E k p K $_{Ptq}$ w ktq

结束于

服务器输出： w_{T}

We now explain how learning to minimize $\mathbf{L}_{NTD}$ preserves global knowledge on out-local distribution p($\mathbf{D}$). Suppose there are N number of data points in the local dataset $\mathbf{D}$. The accumulated Kullback- Leibler divergence loss L_{KL} between $q'f^{l}$, the probability vector for the data x_i, and its reference $q_{g'i}$ to be matched for is:

$$L_{KL} = - N \sum\sum^{N\ C} q^{g,i}(c) \log \frac{Qr^{i}(c)}{qffc)} \tag{12}$$

通过拆分真实类和非真实类的加数，*该术语*变为：

$$L^{true}_{KL} \quad \bar{}_N \sum^{1\ N \ldots} qg\,(yi) \log \frac{q^{l}T^{l}(yi)q}{\wedge(yi)} \qquad n \text{ not-true } ^1 V'' V'' g.i\ 1; \quad L_{KL} = \bar{}\ \sum^{N_C} qg^{(c)} \log \frac{q_{l^{:i}}(c^1)}{q_{g^{:i}}(c_1)} \tag{13}$$

命题 2. *考虑局部分布* p($\mathbf{D}$) = [p 1 .. .p $_C$] *使得* p_c = | SE *及其外部分布* p($\mathbf{D}$) = [p i ,... p c] ， *其中* S_c *是满足* $y_i = c$ *的索引集。那么* $\mathbf{L}_{Ke}$ *和* $\mathbf{L}_{KL}{}^{-true}_{C}$ 每个都等价于 p($\mathbf{D}$) *和* p($\mathbf{D}$) *的加权和为*

$$\begin{aligned} L^{z\text{»true}}_{KL} &\ -y\ p_c \sum_{c=1C} \text{'icSc} \quad -qffc) \log \frac{qr^{i}(c)}{q\wedge' fc)} \\ \frac{L^{z\text{»not-true}}_{KL}}{C-1} &\ -\ p_c \sum_{c=1} {}_{iRSc} \quad -q_{g\cdot i}(c) \log \frac{qr^{i}(c)}{qffc)} \end{aligned} \tag{14}$$

通过等式 13 的少量计算，我们推导出上述命题。推导在附录 Q 中提供。该命题表明，匹配真实类和非真实类 logits 会导致局部分布 p($\mathbf{D}$) 和局部外分布 p($\mathbf{D}$) 的损失。

在我们的 FedNTD （方程 10 ）**的损失函数中，我们通过使用** $\mathbf{L}_{CE}$ 跟踪来自本地数据集中标记数据的真实类信号，获得了关于局部分布的新知识。同时，我们通过遵循全局模型的视角，对应于非真实类信号，使用 $\mathbf{L}_{NTD}$ 保留了关于局部外分布的先前知识。在这里，超参数 f 控制着学习新知识和保留先前知识之间的权衡。这类似于*稳定性-可塑性困境*[39]在持续学习中，学习方法必须平衡保留先前任务中的知识，同时为当前任务学习新知识[35]。

表 1：MNIST [11]、CIFAR-10 [25]、CIFAR-100 [25]和 CINIC-10 [10]的准确度@1 (%) 。括号中的值是忘记度量 *F*。箭头 (t , 1)显示与 FedAvg 的比较。附录 F 中提供了每个实验的标准偏差。

NIID 分区策略：Sharding

方法	MNIST	CIFAR-10 小号 = 2	CIFAR-10 小号 = 3	CIFAR-10 小号 = 5	CIFAR-10 小号 = 10	CIFAR-100	CINIC-10
联邦平均[37]	78.63 (0.20)	40.14 (0.59)	51.10 (0.46)	57.17 (0.37)	64.91 (0.26)	25.57 (0.49)	39.64 (0.59)
联邦曲线[43]	78.56 (0.2i)吨	44.52 (0.53) 1	49.00 (0.47) t	54.61 (0.39) t	62.19 (0.27) t	22.89 (0.49) t	40.45 (0.57) 1
联邦快递[30]	78.26 (0.21)吨	41.48 (0.57) 1	51.65 (0.45) 1	56.88 (0.37) 1	64.65 (0.25) t	25.10 (0.49) 1	41.47 (0.57) 1
联邦新星[47]	77.O4 (0.21) t	42.62 (0.56) 1	52.03 (0.44) 1	62.14 (0.30) 1	66.97 (0.20) 1	26.96 (0.41) 1	42.55 (0.56) 1
脚手架[20]	81.O5 (0.17)	44.60 (0.53) 1	54 . 26 (0.39)	6 **5.7** 4 (0.23)	1 **68.97** (0.16) 1	30.82 (0.36) 1	42.66 (0.54) 1
月亮[28]	76.56 (0.23)吨	38.51 (0.60) t	50.47 (0.47) t	56.69 (0.39) t	65.30 (0.25) 1	25.29 (0.48) t	37.07 (0.62) t
FedNTD （我们的）	**84.44** (0.i3) 1	**52.61** (0.43) 1	**58.18** (0.34) 1	64.93 (0.23) 1	**68.56** (0.15) 1	**31.69** (0.32) 1	**48.O7** (0.48) 1

NIID 分区策略：LDA

方法	MNIST	CIFAR-10 — = 0.05	CIFAR-10 — = 0.1	CIFAR-10 — = 0.3	CIFAR-10 — = *0.5*	CIFAR-100	CINIC-10
联邦平均[37]	7 9. 73 (0.19)	28.24 (0.7i)	46.49 (0.5i	) 57.24 (0.36)	62.53 (0.28)	30.69 (0.32)	38.14 (0.6O)
联邦曲线[43]	78.72 (0.2O) t	33.64 (0.66) 1	44.26 (0.53) t	54.93 (0.38) t	59.37 (0.3O) t	29.16 (0.32) t	36.69 (0.6i) t
联邦快递[30]	79.25 (0.i9) t	37.19 (0.62) 1	47.65 (0.49) 1	57.35 (0.35) 1	62.39 (0.27) t	30.60 (0.32) t	39.47 (0.58) 1
联邦新星[47]	60.37 (0.38) t	10.00 (失败) t	28.06 (0.71) t	57.44 (0.35) 1	64.65 (0.23) 1	32.15 (0.28 1	30.44 (0.68) t
脚手架[20]	71.57 (0.26)	10.00 （生效） t	23.12 (0.74) t	62.O1 (**O.29**)	**66. 16** （0.19） 1	**33.68** (0.25) 1	28.78 (0.69) t
月亮[28]	78.95 (0.20) t	28.35 (0.71) 1	44.77 (0.53) t	58.38 (0.35) 1	63.10 (0.27) 1	30.64 (0.32) t	37.92 (0.6I) t
FedNTD （我们的）	**81 .3** 4 (0.17)	**40 。** 17 (0.58)	**54 . 42** (0.42)	6 **2. 42** (0.29)	**66. 12** （0.19） 1	32.37 (0.26) 1	**46.24** (0.5O) 1

4 实验

4.1 实验装置

我们在 MNIST [11]、CIFAR-10 [25]、CIFAR-100 [25]和 CINIC-10 [10]上测试了我们的算法。我们将数据分发给 100 个客户，并以 0.1 的比例随机抽样客户。对于 CINIC-10，我们使用 200 个客户端，采样率为 0.05。我们使用初始学习率为 0.1 的动量 SGD，动量设置为 0.9。学习率在每一轮都以 0.99 的系数衰减，并应用 1e-5 的权重衰减。我们采用两种不同的 NIID 分区策略：

- **(i) Sharding** [37]：按标签对数据进行排序并将数据划分为相同大小的碎片，并通过 *s*（每个用户的碎片数）控制异质性。在此策略中，仅考虑统计异质性，因为每个客户端的数据集大小相同。我们将 s 设置为 MNIST (s = 2)、CIFAR-10 (se {2, 3,5,10})、CIFAR-100 (s = 10) 和 CINIC-10 (s = 2)。
- **(ii) Latent Dirichlet Allocation (LDA)** [34 , 46]：通过采样*pc* 《Dir(a)分配 c 类的分区。在此策略中，每个客户端的分布和数据集大小都不同。我们将 *a* 设置为 MNIST (a = 0.1)、CIFAR-10 (ae {0.05,0.1,0.3,0.5})、CIFAR-100 (a = 0.1) 和 CINIC-10 (a = 0.1)

附录 B 中提供了有关模型、数据集、超参数和分区策略的更多详细信息。

4.2 数据异构性能

我们将我们的 FedNTD 与各种现有工作进行比较，结果如表 1 所示。正如[27]中所报告的那样，即使是最先进的方法也只能在特定设置中表现良好，并且它们的性能通常会低于 FedAvg。然而，我们的 FedNTD 在所有设置上始终优于基线，在大多数情况下实现了最先进的结果。

对于表 1 中的每个实验，我们还在括号中报告了遗忘 *F* 以及准确度。请注意，较小的 **F** 值表示全局模型较少忘记先前的知识。我们发现联邦学习的性能与遗忘密切相关，随着遗忘的减少而提高。*我们认为，先前作品的收获实际上来自于他们以自己的方式预防遗忘。*

我们强调，从异质本地学习的先前工作通常需要*状态*性（即，客户端应通过识别重复采样）[28、47] 、额外的*通信成本*[20]或*辅助数据*[33] 。然而，我们的 FedNTD 既不会妥协任何潜在的隐私问题，也不会增加额外的沟通负担。附录 C 中提供了简要比较。

我们在附录 G 中进一步对局部时期、客户端采样率和模型架构的影响进行了实验，在附录 H 中进行了非真实蒸馏相对于知识蒸馏的优势，在附录 K 中进行了 FedNTD 超参数的影响。在下一节中，我们将分析 FedNTD 的知识保存如何有利于联邦学习。

5 FedNTD 知识保存

在图 7 中，我们展示了不同异质性水平上的测试准确性。尽管 FedAvg 和 FedNTD 在局部分布 p(x) 上的局部精度几乎没有变化，但 FedNTD 显着提高了局部分布 p(x) 上的局部精度，这意味着它可以防止遗忘。随之而来的是，全局模型的测试精度也大幅提升。当局部 epoch 的数量增加时，这些差距会扩大，局部模型与全局模型的偏差很大。局部训练时的准确率曲线如图 6 所示。这表明快速拟合局部分布

我们感兴趣的是，尽管 FedNTD 在本地分布上的表现几乎没有变化，但 FedNTD 在本地外分布上的知识保存如何对联邦学习有益。为了弄清楚这一点，我们在局部训练后分析了 FedNTD 中的局部模型，并提出了两个主要原因：

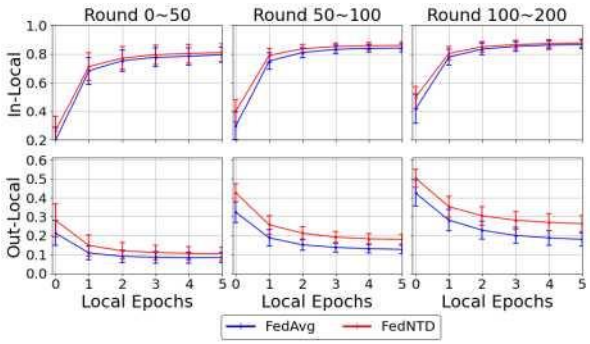


图 6：*本地训练*中 CIFAR-10 (s=2) 的准确度。误差条代表客户端的标准偏差。

- ***重对齐**：每个权重的语义保留了多少？*

leads to forgetting on out-local distribution, But FedNTD effectively relieves this tendency without 损害对本地分布的学习能力。

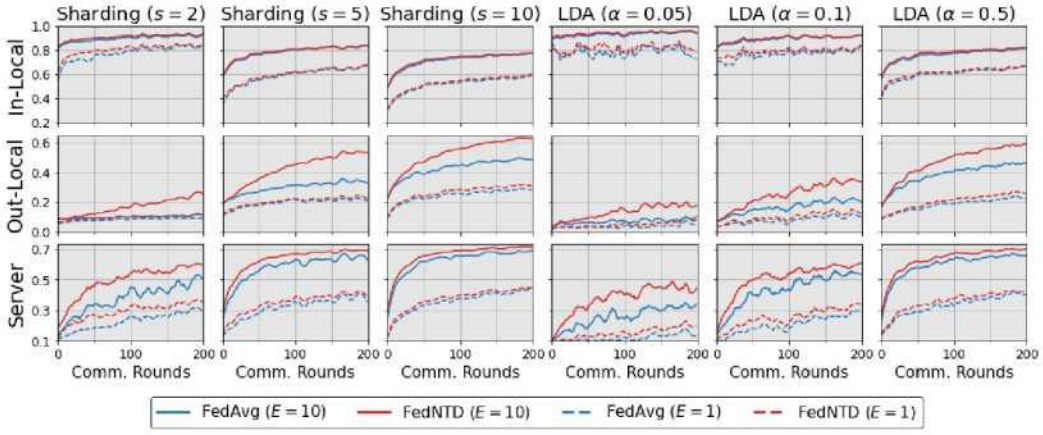


图 7：FedAvg [37]（*蓝线*）和我们的 FedNTD（*红线*）在 CIFAR-10 上的学习曲线，具有针对局部时期 *E* e {1,10} 的各种异质性设置。（第 1 [行] 和第 2 [行]）：本地分布 *p(x)* 和外本地分布 p(x) 的本地测试精度。（第 3 [行]）：全球服务器测试准确性。

- ***重差异**：局部权重偏离全局模型有多远？*

5.1 权重对齐

在最近的研究中，有人提出在局部模型的权重参数中编码的语义信息不匹配，即使对于相同的坐标（即相同的位置）

[46,53,54] 。由于当前的聚合方案平均相同坐标的权重，匹配局部模型之间的语义对齐在全局收敛中起着重要作用。

为了分析每个参数的语义对齐，我们通过平均具有最大激活输出的类来识别单个神经元的类偏好。然后，我们测量两个不同模型之间的层的对齐方式，作为类别偏好彼此匹配的神经元的比例。结果在表 2 中提供。

虽然 FedAvg 和 FedNTD 在 IID 情况下几乎没有差异，但 FedNTD 显着增强了 NIID 情况下的一致性。可视化结果在附录 M 中，其中包含有关如何测量对齐的更多详细信息。我们进一步分析了使用附录 N 中的单位超球体和附录 O 中的 T-SNE 进行局部训练后特征的变化。

表 2：IID 和 NIID 的 CIFAR-10 数据集上分布式全局 (DG)、10 个局部变量 (L) 和聚合全局 (AG) 模型的最后两个 fc 层的对齐（分片 s = 2，LDA a = 0.05）在第 200 轮。

层	结盟	国际身份证		NIID (s 2)		NIID （一个 = 0.05)	
		美联储平均	联邦新台币	美联储平均	F 新台币	美联储平均	联邦新台币
线性_1	\\f¹与 wL¹.q	0.679	0.668	0.635	**0.703**	0.597	**0.756**
（暗淡：512）	w G ¹与 w G ᵗ q	0.850	0.830	0.787	**0.871**	0.670	**0.856**
线性_1	w G ʳ与 W p t q	0.771	0.765	0.488	**0.552**	0.512	**0.730**
（暗淡：128）	w £ ⁻¹)与 w G ᵗq	0.898	0.906	0.609	**0.836**	0.586	**0.859**

5.2 权重差异

FedNTD 的知识保存导致全局模型更均匀地预测每个类别。在这里，我们描述了具有均匀预测性能的全局模型如何稳定权重差异。考虑一个适合特定原始分布的模型，现在它在新分布上进行训练。然后原始模型和拟合模型之间的权重距离随着原始分布和新分布之间的距离的增加而增加。

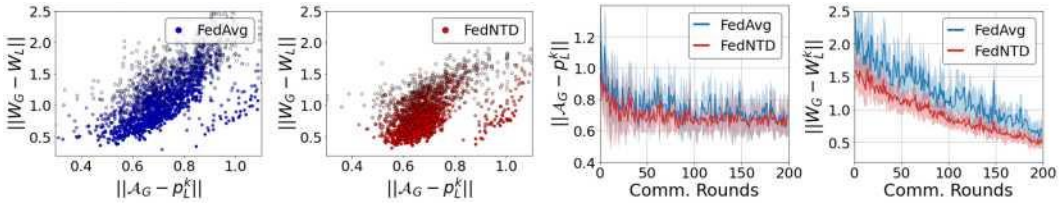


图 8：CIFAR10 (s=2) 上权重和分布的距离。(a), (b)：两个距离之间的关系。后面几轮的不透明度更高。(c)、(d)：200 回合的测量距离。

我们认为，如果全局模型的基础分布与局部分布之间的距离很小，则全局模型与局部模型之间的移动距离也会变近。如果我们假设局部分布是任意生成的，则全局模型的基础分布最稳健的选择是均匀分布。我们正式改写我们的论点如下：

命题 3. 令 P：A 。^ R ￥ ₀ *为客户局部分布的概率测度，* n *为假设中的测度集。假设类损失* $L_c(w) = E_{x|y=c}[\mathbf{L}(x; w)]$ *是* A *平滑的并且* w_c *是* L_c *的最优值。那么分布为* p *的客户损失* L *变为：*

$$L(w) = y\,P\,c\,L\,c\,(w) \wedge y\,P\,c\,L\,c\,(w_c) + \frac{1}{2}\,xVp_c\,\|w - Wc\|^2. \quad (15)$$

立方厘米

然后对于任意 P en *，均匀分布达到极小极大值：*

$$unif.dist\ e\ argmin \sup_{p\,e\,A\,c\,P\,e\,n;} E_{p^{\wedge}p}[\|w_p — w_p\|]，\text{其中 } w_p = p_c\,w_{c。} \quad (16)$$

证明在附录 S 中提供。尽管全局模型的基础分布是未知的，但归一化的类精度向量是它的一个方便的近似值：$\mathbf{A}_{G=A}\blacksquare[a_1,\ldots a_C]$，其中 A 是全局模型的测试精度，a_c 是它在 c 类上的类准确率。

图 8 中的结果从经验上验证了我们的论点。权重差异之间存在很强的相关性$\}w - w_k\|$（对于全局模型 w 和客户端 k 的本地模型 w_k）和分布距离 $\|A_G - p_k\|$（对于客户 k 的分布 p_k）。通过为局部训练提供更好的起点，FedNTD 有效地稳定了权重差异。

6 相关工作

联合学习 (FL)被提议用于更新全局模型，同时本地数据保存在客户端的设备中[23、24] 。标准算法是 FedAvg [37]，它通过平均参数来聚合训练有素的局部模型。尽管其有效性已在 iid 设置[44 , 48]中得到广泛讨论，但当分布式数据是异构的时，许多算法会获得次优[31 , 56]。直到最近，人们提出了 FedAvg 的各种变体来克服这个问题。一线工作侧重于*本地* 通过调整局部模型与全局模型的偏差来进行修改[1 , 20 , 30 , 47]。另一个是*服务器端修改*，它提高了服务器中本地模型聚合的效率[9,33,46,55] 。我们的工作旨在在本地培训期间保留全球知识，这属于本地方面的方法。

Forgetting View in FL 考虑在 FL 中遗忘的先驱工作是 FedCurv [43]。它将每个本地客户端视为一个任务，并且[43]调节本地参数的变化以防止所有其他客户端的准确性下降。但是，它需要跨客户端计算和传达参数方面的重要性，这严重加重了学习过程的负担。另一方面，我们关注类遗忘，并建议来自本地数据的非真实逻辑包含足够的知识来防止它。我们研究的一项并行工作是[49]，它还通过经验表明在本地培训后先前学习的数据损失增加，从而报告了本地客户的遗忘问题。为了防止遗忘，[49]利用生成的伪数据。相反，我们专注于类遗忘，并建议来自本地数据的非真实逻辑包含足够的知识来防止它。附录 D 中进一步讨论了持续学习文献。

FL 中的**知识蒸馏 (KD) 在** FL 中，一种典型的方法是使用 KD 使全局模型从局部模型的集合中学习[9、26、33、57] 。通过利用未标记的辅助数据，KD 通过丰富聚合有效地解决了局部漂移。然而，这种精心设计的代理数据可能并不总是可用[30、55、58] 。尽管最近的作品生成伪数据以通过无数据 KD [55、58]提取知识，它们需要额外的繁重计算，并且样本的质量对过程中涉及的许多超参数很敏感。另一方面，作为 KD 的简单变体，我们提出的方法在没有任何额外资源要求的情况下在异质性场景中表现出色。

7 结论

这项研究从类比到持续学习开始，并表明遗忘可能是联邦学习中的一个主要问题。我们的观察表明，局部分布之外的知识在局部训练中容易被遗忘，并且与不稳定的全局收敛密切相关。为了克服这个问题，我们提出了一种简单而有效的算法 FedNTD，它只对非真实类进行局部蒸馏以防止遗忘。与以前的方法不同，FedNTD 没有任何额外要求。我们从多个角度分析了 FedNTD 的效果，并展示了它在联邦学习中的好处。

更广泛的影响我们认为联邦学习是一种重要的学习范式，可以实现隐私保护 ML。我们的工作提出了遗忘问题，并介绍了在不损害数据隐私的情况下重温它的方法。这项工作背后的洞察力可能会激发新的研究。然而，所提出的方法在全局模型中维护局部分布之外的知识。这意味着如果全局模型有偏差，则经过训练的局部模型更容易出现类似的趋势。ML 参与者应该考虑这一点。

致谢

这项工作得到了韩国政府 (MSIT) 资助的信息与通信技术规划与评估研究所 (IITP) 赠款的支持 [No. 2021-0-00907, 实现主动即时响应和快速学习的自适应轻量级边缘协同分析技术的开发, 90%] [No. 2019-0-00075, 人工智能研究生院项目(KAIST), 10%].

参考

[1] Durmus Alp Emre Acar、Yue Zhao、Ramon Matas Navarro、Matthew Mattina、Paul N Whatmough 和 Venkatesh Saligrama。基于动态正则化的联邦学习。*arXiv 预印本 arXiv:2111.04263* ,2021。

[2] Mohammed Aledhari、Rehma Razzak、Reza M Parizi 和 Fahad Saeed。联邦学习：关于支持技术、协议和应用程序的调查。*IEEE 访问*，8:140699-140725，2020 年。

[3] Rahaf Aljundi、Francesca Babiloni、Mohamed Elhoseiny、Marcus Rohrbach 和 Tinne Tuytelaars。记忆感知突触：学习（不）忘记什么。在*欧洲计算机视觉会议 (ECCV) 的会议记录中*，第 139-154 页，2018 年。

[4] Sungmin Cha、Hsiang Hsu、Taebaek Hwang、Flavio P Calmon 和 Taesup Moon。Cpr：用于持续学习的分类器投影正则化。*arXiv 预印本 arXiv:2006.07326，2020* 年。

[5] Arslan Chaudhry、Puneet K Dokania、Thalaiyasingam Ajanthan 和 Philip HS Torr。增量学习的黎曼步走：理解遗忘和不妥协。在*欧洲计算机视觉会议 (ECCV) 会议记录中*，第 532-547 页，2018 年。

[6] Arslan Chaudhry、Albert Gordo、Puneet Kumar Dokania、Philip Torr 和 David Lopez-Paz。使用后见之明将过去的知识锚定在持续学习中。*arXiv 预印本 arXiv:2002.08165* , 2(7), 2020。

[7] Arslan Chaudhry、Naeemullah Khan、Puneet K Dokania 和 Philip HS Torr。在低秩正交子空间中持续学习。*arXiv 预印本 arXiv:2010.11635，2020* 年。

[8] Arslan Chaudhry、Marc'Aurelio Ranzato、Marcus Rohrbach 和 Mohamed Elhoseiny。使用 a-gem 进行高效的终身学习。*arXiv 预印本 arXiv:1812.00420，2018* 年。

[9] 陈红友、赵伟伦．Fedbe：使贝叶斯模型集成适用于联邦学习。*arXiv 预印本 arXiv:2009.01974，2020* 年。

[10] Luke N Darlow、Elliot J Crowley、Antreas Antoniou 和 Amos J Storkey。Cinic-10 不是 imagenet 或 cifar-10。*arXiv 预印本 arXiv:1810.03505，2018* 年。

[11] 李登. 用于机器学习研究的手写数字图像的 mnist 数据库。*IEEE 信号处理杂志*, 29(6):141-142, 2012。

[12] 特伦斯·德弗里斯和格雷厄姆·W·泰勒。改进了带切口的卷积神经网络的正则化。*arXiv 预印本 arXiv:1708.04552，2017* 年。

[13] 董家华, 王立旭, 方振, 孙甘, 许世超, 王晓, 朱其. 联邦类增量学习。*arXiv 预印本 arXiv:2203.11473，2022* 年。

[14] Mehrdad Farajtabar、Navid Azizan、Alex Mott 和李安。用于持续学习的正交梯度下降。在*人工智能和统计国际会议上*，第 3762-3773 页。2020 年 PMLR。

[15] Ian J Goodfellow、Mehdi Mirza、Da Xiao、Aaron Courville 和 Yoshua Bengio。基于梯度的神经网络中灾难性遗忘的实证研究。*arXiv 预印本 arXiv:1312.6211，2013* 年。

[16] 何朝阳, 李松泽, 苏金贤, 曾晓, 张咪, 王宏义, 王晓阳, Praneeth Vepakomma, Abhishek Singh, 秋航, 等。Fedml：联邦机器学习的研究库和基准。*arXiv 预印本 arXiv:2007.13518，2020* 年。

[17] Geoffrey Hinton、Oriol Vinyals 和 Jeff Dean。在神经网络中提炼知识。*arXiv 预印本 arXiv:1503.02531，2015* 年。

[18] Tzu-Ming Harry Hsu、Hang Qi 和 Matthew Brown。测量非相同数据分布对联合视觉分类的影响。*arXiv 预印本 arXiv:1909.06335，2019* 年。

[19] Peter Kairouz、H Brendan McMahan、Brendan Avent、Aurdlien Bellet、Mehdi Bennis、Ar-jun Nitin Bhagoji、Keith Bonawitz、Zachary Charles、Graham Cormode、Rachel Cummings 等。联邦学习的进展和未解决的问题。*arXiv 预印本 arXiv:1912.04977，2019* 年。

[20] Sai Praneeth Karimireddy、Satyen Kale、Mehryar Mohri、Sashank Reddi、Sebastian Stich 和 Ananda Theertha Suresh。支架：联邦学习的随机控制平均。在*机器学习国际会议上*，第 5132-5143 页。2020 年 PMLR。

[21] Ahmed Khaled、Konstantin Mishchenko 和 Peter Richtdrik。针对相同和异构数据的本地 sgd 的更严格理论。在*国际人工智能与统计会议上*，第 4519-4529 页。2020 年 PMLR。

[22] James Kirkpatrick、Razvan Pascanu、Neil Rabinowitz、Joel Veness、Guillaume Desjardins、Andrei A Rusu、Kieran Milan、John Quan、Tiago Ramalho、Agnieszka Grabska-Barwinska 等。克服神经网络中的灾难性遗忘。*美国科学院院刊*, 114(13):3521-3526, 2017.

[23] Jakub Konecny、H Brendan McMahan、Daniel Ramage 和 Peter Richtarik。联合优化：用于设备智能的分布式机器学习。*arXiv 预印本 arXiv:1610.02527，2016* 年。

[24] Jakub Konecny、H Brendan McMahan、Felix X Yu、Peter Richtarik、Ananda Theertha Suresh 和 Dave Bacon。联邦学习：提高沟通效率的策略。*arXiv 预印本 arXiv:1610.05492，2016* 年。

[25] Alex Krizhevsky、Vinod Nair 和 Geoffrey Hinton。Cifar-10 和 cifar-100 数据集。*网址：https://www. CS。多伦多。edu/kriz/cifar。html* , 6, 2009.

[26] 李大良，王军普．Fedmd：通过模型蒸馏实现异构联邦学习。*arXiv 预印本 arXiv.1910.03581，2019* 年。

[27] 李勤斌, 刁益群, 陈权, 何炳生. 非独立同分布数据孤岛的联邦学习：一项实验研究。*arXiv 预印本 arXiv:2102.02079，2021* 年。

[28] 李勤斌, 何炳生, 宋道明. 模型对比联邦学习。在 *IEEE/CVF 计算机视觉和模式识别会议记录中*，第 10713-10722 页，2021 年。

[29] 李田、阿尼特 痄马尔 萨胡、阿米特 塔尔沃卡和弗吉尼亚 史密斯。联邦学习：挑战、方法和未来方向。*IEEE 信号处理杂志*，37(3):50-60, 2020。

[30] 田力、Anit Kumar Sahu、Manzil Zaheer、Maziar Sanjabi、Ameet Talwalkar 和 Virginia Smith。异构网络中的联合优化。*机器学习和系统会议记录*，2:429-450，2020 年。

[31] 李翔, 黄凯旋, 杨文浩, 王树森, 张志华. 关于 fedavg 在非 iid 数据上的收敛。*arXiv 预印本 arXiv:1907.02189，2019* 年。

[32] 李志忠和 Derek Hoiem. 学而不忘。*IEEE 交易模式分析和机器智能*，40(12):2935-2947, 2017。

[33] 陶林、孔令靖、塞巴斯蒂安 乌 斯蒂奇和马丁 贾吉。联邦学习中鲁棒模型融合的集成蒸馏。*arXiv 预印本 arXiv:2006.07242，2020* 年。

[34] 米罗, 陈飞, 胡大鹏, 张一凡, 梁健, 冯嘉世. 不用担心异质性：使用非独立同分布数据进行联邦学习的分类器校准。*arXiv 预印本 arXiv:2106.05001，2021* 年。

[35] Marc Masana、Xiali Liu、Bartlomiej Twardowski、Mikel Menta、Andrew D Bagdanov 和 Joost van de Weijer。类增量学习：图像分类的调查和性能评估。*arXiv 预印本 arXiv:2010.15277，2020* 年。

[36] 迈克尔 麦克洛斯基和尼尔 J 科恩。连接主义网络中的灾难性干扰：顺序学习问题。在*学习和动机心理学中*，第 24 卷，第 109-165 页。爱思唯尔，1989 年。

[37] Brendan McMahan、Eider Moore、Daniel Ramage、Seth Hampson 和 Blaise Aguera y Arcas。从分散数据中高效地学习深度网络的通信。在*人工智能和统计*中，第 1273-1282 页。人民解放军，2017 年。

[38] Matias Mendieta, Taojiannan Yang, Pu Wang, Minwoo Lee, Zhengming Ding, and Chen Chen. 本地学习很重要：重新思考联邦学习中的数据异质性。在 *IEEE/CVF 计算机视觉和模式识别会议记录中*，第 8397-8406 页，2022 年。

[39] Martial Mermillod、Aurelia Bugaiska 和 Patrick Bonin。稳定性-可塑性困境：研究从灾难性遗忘到年龄限制学习效果的连续统一体。*心理学前沿*，4:504，2013 年。

[40] German I Parisi、Ronald Kemker、Jose L Part、Christopher Kanan 和 Stefan Wermter。使用神经网络进行持续终身学习：综述。*神经网络*, 113:54-71, 2019.

[41] Adam Paszke、Sam Gross、Francisco Massa、Adam Lerer、James Bradbury、Gregory Chanan、Trevor Killeen、Zeming Lin、Natalia Gimelshein、Luca Antiga、Alban Desmaison、Andreas Kopf、Edward Yang、Zachary DeVito、Martin Raison、Alykhan Tejani 、Sasank Chilamkurthy、Benoit Steiner、Lu Fang、Junjie Bai 和 Soumith Chintala。Pytorch：一种命令式、高性能的深度学习库。由 H. Wallach、H. Larochelle、A. Beygelzimer、F. d'Alche-Buc、E. Fox 和 R. Garnett 编辑，*神经信息处理系统进展 32*，第 8024-8035 页。柯伦联合公司，2019 年。

[42] 马克 B 环。儿童：迈向持续学习的第一步。在*学习学习中*，第 261-292 页。施普林格，1998 年。

[43] Neta Shoham、Tomer Avidor、Aviv Keren、Nadav Israel、Daniel Benditkis、Liron Mor-Yosef 和 Itai Zeitak。克服非独立同分布数据联邦学习中的遗忘。*arXiv 预印本 arXiv:1910.07796，2019* 年。

[44] 塞巴斯蒂安 斯蒂奇。本地 sgd 收敛快，通信少。*arXiv 预印本 arXiv:1805.09767，2018* 年。

[45] 塞巴斯蒂安 特伦。终身学习算法。在*学习学习中*，第 181-209 页。施普林格，1998 年。

[46] Hongyi Wang、Mikhail Yurochkin、Yuekai Sun、Dimitris Papailiopoulos 和 Yasaman Khazaeni。具有匹配平均的联邦学习。*arXiv 预印本 arXiv:2002.06440，2020* 年。

[47] 王建宇、刘庆华、梁浩、Gauri Joshi 和 H Vincent Poor。解决异构联邦优化中的目标不一致问题。*arXiv 预印本 arXiv:2007.07481，2020* 年。

[48] Blake Woodworth、Kumar Kshitij Patel、Sebastian Stich、Zhen Dai、Brian Bullins、Brendan Mcmahan、Ohad Shamir 和 Nathan Srebro。local sgd 比 minibatch sgd 好吗？在*机器学习国际会议上*，第 10334-10343 页。2020 年 PMLR。

[49] 徐晨城, 洪志伟, 黄敏烈, 姜涛. 通过减轻局部训练中的遗忘加速联邦学习。*arXiv 预印本 arXiv:2203.02645，2022* 年。

[50] 杨强，刘洋，陈天健，童永新. 联合机器学习：概念和应用。*ACM 智能系统和技术交易 (TIST)*，10(2):1-19，2019。

[51] Jaehong Yoon、Wonyong Jeong、Giwoong Lee、Eunho Yang 和 Sung Ju Hwang。联合持续学习与加权客户端间迁移。在*机器学习国际会议上*，第 12073-12086 页。PMLR，2021 年。

[52] Tehrim Yoon、Sumin Shin、Sung Ju Hwang 和 Eunho Yang。Fedmix：平均增强联邦学习下的混合近似。在*国际学习代表大会上*，2021 年。

[53] 于付勋, 张维山, 秦竹伟, 徐子瑞, 王迪, 刘晨晨, 田志, 陈向. Fed2：特征对齐的联邦学习。在*第 27 届 ACM SIGKDD 知识发现与数据挖掘会议论文集中*，第 2066-2074 页，2021 年。

[54] Mikhail Yurochkin 、 Mayank Agarwal 、 Soumya Ghosh 、 Kristjan Greenewald 、 Nghia Hoang 和 Yasaman Khazaeni。神经网络的贝叶斯非参数联邦学习。在*机器学习国际会议上*，第 7252-7261 页。人民解放军，2019 年。

[55] 张琳，沉力，丁亮，陶大成，段玲玉。通过非独立同分布联邦学习的无数据知识蒸馏微调全局模型。*arXiv 预印本 arXiv:2203.09249，2022* 年。

[56] Yue Zhao、Meng Li、Liangzhen Lai、Naveen Suda、Damon Civin 和 Vikas Chandra。非独立同分布数据的联邦学习。*arXiv 预印本 arXiv:1806.00582，2018* 年。

[57] 周延林, George Pu, 马希尧, 李小林, 吴大鹏. 蒸馏的一次性联邦学习。*arXiv 预印本 arXiv:2009.07999，2020* 年。

[58] 朱庄迪, 洪俊元, 周嘉宇. 异构联邦学习的无数据知识蒸馏。*arXiv 预印本 arXiv:2105.10056，2021* 年。

检查清单

清单遵循参考文献。请仔细阅读清单指南，了解如何回答这些问题的信息。对于每个问题，将默认**[TODO]**更改为[Yes] 、[No]或[N/A] 。强烈建议您通过引用论文的适当部分或提供简短的内联描述**来为您的答案提供理由。**例如：

- 您是否包括代码和数据集的许可？[Yes]数据集为公开数据集，代码为 MIT license

请不要修改问题，仅使用提供的宏作为您的答案。请注意，清单部分不计入页数限制。在你的论文中，请删除这个说明块，只保留上面的清单部分标题以及下面的问题/答案。

1. 对于所有作者...
   (a) 摘要和引言中的主要主张是否准确反映了论文的贡献和范围？[是]我们在摘要和引言中明确了范围。第 1 节末尾总结了这些贡献。
   (b) 你有没有描述你工作的局限性？[是]在参考之前，我们讨论了作为广泛影响的潜在限制
   (c) 您是否讨论过您的工作对社会的任何潜在负面影响？[是]在参考之前，我们讨论了更广泛的影响
   (d) 您是否已阅读伦理审查指南并确保您的论文符合这些指南？[是的]
2. 如果你包括理论结果.....
   (a) 你是否陈述了所有理论结果的全套假设？[是]在所有命题部分。详细的符号在附录 A 中。
   (b) 你是否包括了所有理论结果的完整证明？[是]包含在相应的附录部分中。
3. 如果你进行实验.....
   (a) 您是否包括了重现主要实验结果所需的代码、数据和说明（在补充材料中或作为 URL）？[是]包含在补充材料中
   (b) 您是否指定了所有训练细节（例如，数据拆分、超参数、它们是如何选择的）？[是]包含在附录中。
   (c) 您是否报告了错误条（例如，关于多次运行实验后的随机种子）？[是]我们在相应的附录部分中指出了主要实验的每个结果的标准偏差。
   (d) 您是否包括计算总量和使用的资源类型（例如，GPU 类型、内部集群或云提供商）？[是]包含在附录中。
   (e) 如果您的作品使用现有资产，您是否引用了创作者？[是的]
   (f) 你有没有提到资产的许可证？[是的]
   (g) 您是否在补充材料中或作为 URL 包含了任何新资产？[是的]
   (h) 您是否讨论过是否以及如何获得您正在使用/管理其数据的人的同意？[N/A]我们使用了公共基准。
   (i) 您是否讨论过您正在使用/管理的数据是否包含个人身份信息或攻击性内容？[N/A]我们只使用了已充分考虑问题的公共数据集。
4. 如果你使用众包或对人类受试者进行研究.....
   (a) 如果适用，您是否提供了给参与者的说明的全文和屏幕截图？[不适用]
   (b) 您是否描述了任何潜在的参与者风险，并提供机构审查委员会 (IRB) 批准的链接（如果适用）？[不适用]
   (c) 您是否包括支付给参与者的预估小时工资和用于参与者补偿的总金额？[不适用]

符号表

表 3：整篇论文的符号表。

	类索引 (ce{1,..., *C}* = *[C*])
	数据索引（即 {1, ■ ■■, *N* } = [N]）
指数：	客户指数 (k e {1, ■ ■■, *K* } = *[K*] 或 e $K^{(i)}$ *)*
C , *C i k; k ·t*	回合索引 （t e {1, ■■■, *T* } = [**T**]）
电子	局部时期的索引 (te {1, ■■■, *E}* = [E])
参数：	
秒	狭利克雷分布的参数
p	每个用户的分片数量
	蒸馏损失的超参数；一般控制发散损失的相对权重
	softmax 上的温度
7	学习率
数据和权重：**D** D^{k}	
X	整个数据集
是	本地数据集
瓦（吨）	基准
||WG - W_{t} \\	基准的类标签
	服务器模型在第 t 轮的权重
	第 k 个客户端模型在第 t 轮的权重
	在所有轮次中，服务器和客户端模型之间的 L 1 范数的集合.
Softmax 概率：	
q_{T} / q s q^{g} /	teacher/student 模型的软化 softmax 概率
$q \sim_{g}$ / q^{l}	服务器/客户端模型上的 softmax 概率
	在服务器/客户端模型上没有真正类 logit 的情况下计算的软化 softmax 概率
数据集的类别分布：	
	数据集的本地分布
p = [P	数据集的外地分布
损失函数：	
LCE _	
LKL_	
有限公司_	交叉熵损失 Kullback-Leibler 散度 建议的非真实蒸馏损失
准确性和遗忘措施：	
一个(吨)	
*C*C*	
}A G - P L \	服务器模型在第 c 类上的准确性，在第 t 轮。在所有轮次中，在每个客户端上收集*归一化*全局准确度和数据分布之间的 L 1 范数. 向后传输 (BwT)。对于联邦学习情况，我们在服务器模型上计算这个度量。

B 实验设置

在这里，我们提供了实验设置的详细信息。代码由 PyTorch [41]实现，整体代码结构基于 FedML [16]库并进行了一些修改。我们使用 1 个 Titan-RTX 和 1 个 RTX 2080Ti GPU 卡。论文实验中没有进行多 GPU 训练。

8.1 模型架构

我们实验中使用的模型架构来自[37]，它由两个卷积层、最大池化层和两个全连接层组成。[28 , 34]中使用了类似的架构。

8.2 数据集

我们主要使用了四个基准：MNIST [11]、CIFAR-10 [25]、CIFAR-100 [25]和 CINIC-10 [10]。表 4 中描述了有关每个数据集和设置的详细信息。我们使用随机裁剪、水平翻转和归一化来扩充训练数据。对于 MNIST、CIFAR-10 和 CIFAR-100，我们添加了 Cutout [12]增强。

表 4：实验中使用的详细数据集设置。

数据集	**MNIST**	**CIFAR-10**	**CIFAR-100**	**CINIC-10**
数据集类	10	10	100	10
数据集大小	50,000	50,000	50,000	90,000
客户数量	100	100	100	200
客户端采样率	0.1	0.1	0.1	0.05
本地时代 (E)	3 个	5 个	5 个	5 个
批量大小 (B)	50	50	50	50

8.3 学习设置

我们使用初始学习率为 0.01 的动量 SGD 优化器，动量设置为 0.9。动量仅用于本地训练，这意味着动量信息不会传送到服务器。学习率在每一轮以 0.99 的系数衰减，并应用 0.00001 的权重衰减。在第 3 节的动机实验中，我们将学习率固定为 0.01。由于我们假设一个同步联邦学习场景，并行分布式学习是通过顺序训练采样客户端然后将它们聚合为全局模型来模拟的。

8.4 算法实现细节

对于实现的算法，我们搜索超参数并从候选者中选择最好的。表 5 中列出了每种算法的超参数。

表 5：实验中使用的特定于算法的超参数。

方法	**超参数搜索候选**		
联邦平均[37]	没有任何		
联邦曲线[43]	*s* = 500, A = 1.0	没有任何	{0.1, 0.5, 1.0}
联邦快递[30]	^ = 0.1		
联邦新星[47]	没有任何	s *e* {250, 500}, A e ^ e {0.1, 0.5,1.0}	无
脚手架[20]	没有任何	没有任何	
月亮[28]	м = 1.0, т = *0.5*	^ e {0.1, 0.5,1.0},	т e {0.1, 0.5,1, 0}
FedNTD（我们的）	0=1.0，т =1.0	没有任何	

8.5 非独立同分布划分策略

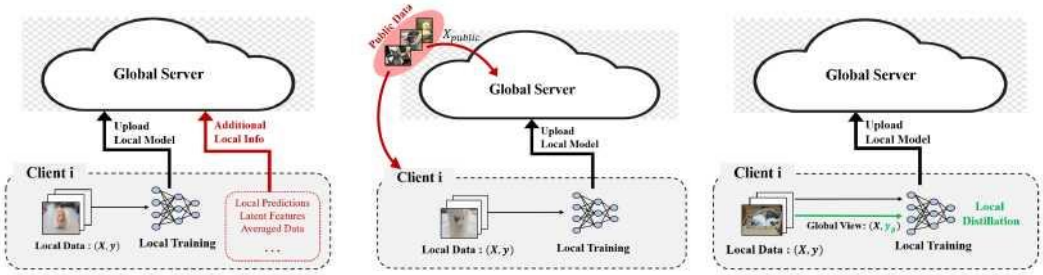
为了广泛解决异构联邦学习场景，我们使用以下两种策略将数据分发到本地客户端：（1） **分片**和（2） **潜在狄利克雷分配**（LDA）。

Sharding 在 Sharding 策略中，我们将数据按标签排序，并将其划分为相同大小的分片，而不会重叠。详细地说，一个分片包含相同类别样本的 n_{DS} 大小，其中 D 是总数据集大小，N 是客户端总数，s 是每个用户的分片数。然后，我们将 s 个分片分配给每个客户端。s 控制本地数据分布的异构性。异质性级别随着每个用户的分片变小而增加，反之亦然。请注意，我们仅测试分片策略中的统计异质性（类分布在本地客户端的偏度），并且本地数据集的大小是相同的。

Latent Dirichlet Allocation (LDA)在 LDA 策略中，每个客户端 k 都分配有 $p_{c,k}$*类 c 训练样本的比例，*其中 $p_c \sim \mathrm{Dir}_K(a)$和 a 是控制异质性的浓度参数。异质性水平随着浓度参数 a 变小而增加，反之亦然。请注意，在 LDA 策略中，类分布和本地数据集大小在本地客户端之间是不同的。

C 与先前作品的概念比较

联合蒸馏方法的概念图如图 9 所示。现有算法要么使用额外的本地信息（图 9a ），要么需要辅助（或*代理*）数据来进行蒸馏。另一方面，我们提出的 FedNTD 没有这样的限制（图 9c ）。



(a) 使用附加本地信息　　(b) 使用辅助数据集　　(c) FedNTD（我们

图 9：联合蒸馏方法概述。

表 6：与 FedAvg 相比的额外资源需求。

方法	无附加要求：		
	有状态？	沟通成本？	辅助数据？
联邦合奏[33]	4 个	4 个	X
联邦调查局[9]	4 个	X	X
月亮[28]	X	4 个	4 个
脚手架[20]	X	X	4 个
FedNTD（我们的）	4 个	4 个	4 个

D 相关工作：持续学习

持续学习 (CL) 是一种学习范式，它更新一系列任务，而不是立即对整个数据集进行训练[42、45] 。在 CL 中，主要挑战是避免灾难性遗忘[15]，即新任务的训练会干扰之前的任务。现有方法试图通过各种策略来缓解这个问题。在*基于参数的*方法中，测量参数对前一个任务的重要性以限制它们的变化[3 , 5 , 22]。*基于正则化的*方法[4 , 32]引入正则化项以防止遗忘。*基于记忆的*方法[6、8]从之前的任务中保留一小段情景记忆并重放它以保持知识。我们的工作更多地与基于正则化的方法相关，引入了额外的本地客观术语以防止忘记本地知识。

值得一提的是，有些作品试图在联邦学习设置中解决经典的持续学习问题。例如，[51]研究了每个本地客户端必须学习一系列任务的场景。这里，特定于任务的参数从全局参数中分解出来，以最小化任务之间的干扰。在[13]中，旧类的关系知识通过类感知梯度补偿逐轮传输。

E 服务器预测一致性

我们将第 3.1 节中的动机实验扩展到主要实验设置。在这里，我们绘制了每个案例的*归一化*类别测试准确度，以确定每个类别对当前准确度的贡献。这有助于观察预测一致性，而不管非 IID 情况下全球服务器准确性的高度波动如何。如图 10 所示， FedNTD 有效地保留了前几轮的知识；因此，全局服务器模型变得比 FedAvg 更早地平均预测每个类。请注意，我们将逐轮测试精度归一化为逐轮方式，这使得每轮的总和为 1.0。

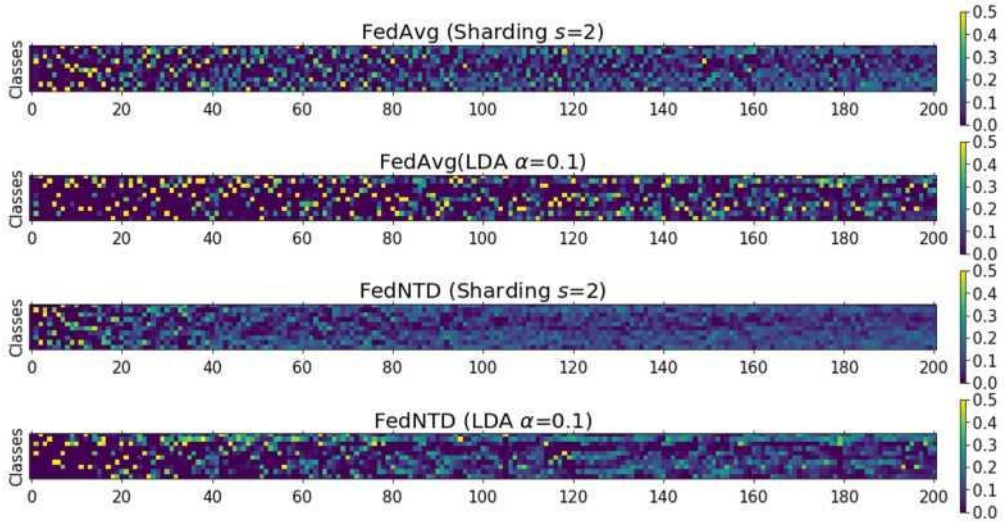


图 10：FedAvg 和 FedNTD 在 CIFAR-10 数据集上的可视化服务器测试准确性。

F 带标准差的实验表

表 7：MNIST [11]、CIFAR-10 [25]、CIFAR-100 [25]和 CINIC-10 [10]的准确度@1 (%) 。括号中的值是标准偏差。箭头 (1 , 1)显示与 FedAvg 的比较。

NIID 分区策略：Sharding							
方法	**MNIST**	s=2	**CIFAR-10** 小号 = 3	小号= 5	小号 = 10	**CIFAR-100**	**CINIC-10**
联邦平均[37]	78.63 +0.42	40.14 +1.15	51.10 +0.11	57.17 +0.12	64.91 +0.69	25.57 +0.44	39.64 +0.78
联邦曲线[43]	78.56 +0.23 1	44.52 +0.44 1	49.00 +0.41	1 54.61 +0.20 1	62.19 +0.47 1	22.89 +0.66 1	40.45 +0.25 1
联邦快递[30]	78.26 +0.28 1	41.48 +1.08 1	51.65 +0.53	1 56.88 +0.15 1	64.65 +0.61 1	25.10 +0.67 1	41.47 +0.99 1
联邦新星[47]	77.04 +0.98 1	42.62 +1.32 1	52.03 +1.49	1 62.14 +0.74 1	66.97 +0.39 1	26.96 +0.59 1	42.55 +0.10 1
脚手架[20]	81.05 +0.26 1	44.60 +2.24 1	54.26 +0.22	1 **65.74** +0.36 1	**68.97** +0.34 1	30.82 +0.31 1	42.66 +0.92 1
月亮[28]	76.56 +0.24 1	38.51 +0.96 1	50.47 +0.15	1 56.69 +0.11 1	65.30 +0.51 1	25.29 +0.24 1	37.07 +0.24 1
FedNTD（我们的）	**84.44** +0.43 1	**52.61** +1.00 1	**58.18** +1.42	1 64.93 +0.34 1	68.56 +0.24 1	**31.69** +0.13 1	**48.07** +0.36 1

NIID 分区策略：LDA							
方法	**MNIST**	一 = 0.05	**CIFAR-10** 一 = 0.1	一 = 0.3	一 = 0.5	**CIFAR-100**	**CINIC-10**
联邦平均[37]	79.73 +0.20	28.24 +3.11	46.49 +0.93	57.24 +0.21	62.53 +0.41	30.69 +0.27	38.14 +3.40
联邦曲线[43]	78.72 +0.44 1	33.64 +2.98 1	44.26 +0.79	1 54.93 +0.46 1	59.37 +0.24 1	29.16 +0.22 1	36.69 +3.03 1
联邦快递[30]	79.25 +0.16 1	37.19 +3.17 1	47.65 +0.90	1 57.35 +0.40 1	62.39 +0.31 1	30.60 +0.16 1	39.47 +3.40 1
联邦新星[47]	60.37 +2.71 1	10.00 （失败）1	28.06 +0.12	1 57.44 +1.69 1	64.65 +0.34 1	32.15 + 0.13 1	30.44 +1.35 1
脚手架[20]	71.57 +0.72 1	10.00 （失败）1	23.12 +0.55	1 **62.01** +0.34 1	**66.16** +0.13 1	33.68 +0.13 1	28.78 +1.26 1
月亮[28]	78.95 +0.46 1	28.35 +3.68 1	44.77 +1.12	1 58.38 +0.09 1	63.10 +0.00 1	30.64 +0.12 1	37.92 +3.31 1
FedNTD（我们的）	81.34 **+0.33** 1	**40.17** +3.19 1	**54.42** +0.06	1 **62.42** +0.53 1	**66.12** +0.26 1	32.37 +0.02 1	**46.24** +1.67 1

G 附加实验

G.1 局部时期的影响

表 8：CIFAR-10 上的 Accuracy@1 （分片 $s = 2$）。括号中的值为遗忘 F。

NIID 分区策略：Sharding (s = 2)					
方法	**本地时代 (E)** 1 个	3 个	5 个	10	20
联邦平均[37]	29.49 (0.70)	34.49 (0.64)	40.14(0.59)	50.08 (0.49)	56.93 (0.42)
联邦快递[30]	29.44 (0.69) 1	34.00 (0.64) 1	41.48 (0.57) 1	42.74 (0.53) 1	43.39 (0.52) 1
联邦新星[47]	27.77 (0.71) 1	32.00 (0.64) 1	42.62 (0.56) 1	48.59 (0.50) 1	58.24 (0.39) 1
脚手架[20]	34.46 (0.64) 1	39.26 (0.58) 1	44.60 (0.53) 1	55.35 (0.41) 1	**62.80** (0.34) 1
FedNTD（我们的）	**35.77** (0.64) 1	**45.47** (0.50) 1	**52.61** (0.43) 1	**60.22** (0.36) 1	60.61 (0.34) 1

表 9：CIFAR-10 上的 Accuracy@1 (LDA $a = 0.1$)。括号中的值为遗忘 F。

NIID 分区策略：LDA (a = 0.1)					
方法	**本地时代 (E)** 1 个	3 个	5 个	10	20
联邦平均[37]	29.77 (0.69)	37.70 (0.60)	46.49 (0.51)	53.80 (0.43)	57.70 (0.39)
联邦快递[30]	33.37 (0.65) 1	37.88 (0.57) 1	47.65 (0.49) 1	44.02 (0.50) 1	44.98 (0.49) 1
联邦新星[47]	26.35 (0.73) 1	24.37 (0.74) 1	28.06 (0.71) 1	47.41 (0.50) 1	10.00（失败）1
脚手架[20]	13.36 (0.86) 1	22.04 (0.75) 1	23.12(0.74) 1	38.49 (0.57) 1	47.07 (0.47) 1
FedNTD（我们的）	**33.94** (0.64) 1	**45.92** (0.50) 1	**54.42** (0.42) 1	**60.67** (0.33) 1	**62.25** (0.30) 1

G.2 采样率的影响

表 10：精度©！在 CIFAR-10 上（分片 $s = 2$）。括号中的值为遗忘 F。

方法	NIID 分区策略：Sharding (s = 2) 客户端采样率 (R) 0.05	0.1	0.3	0.5	1.0
联邦平均[37]	33.06 (0.66)	40.14 (0.59)	49.99 (0.46)	52.98 (0.41)	51.48 (0.30)
联邦快递[30]	35.36 (0.63)女	41.48 (0.57)女	44.54 (0.45) 1	50.02 (0.31) 1	52.53 (0.06)女
联邦新星[47]	29.99 (0.69) 1	42.62 (0.56) 女	55.59 (0.31)女	56.75 (0.23)女	51.89 (0.34) 女
脚手架[20]	29.15 (0.70) 1	44.60 (0.53)女	55.59 (0.31)女	56.75 (0.23)女	57.88 (0.10) 女
FedNTD（我们的）	**46.99** (0.51)女	**52.61** (0.43)女	**59.37** (0.28)女	**60.70** (0.18)楼	**61.53** (0.04)女

表 11：精度©！在 CIFAR-10 上（LDA $a = 0.1$）。括号中的值为遗忘 F。

方法	NIID 分区策略：LDA (a = 0.1) 客户端采样率 (R) 0.05	0.1	0.3	0.5	1.0
联邦平均[37]	29.35 (0.70)	46.49 (0.51)	53.73 (0.39)	58.72 (0.25)	61.38 (0.04)
联邦快递[30]	36.36 (0.63) 女	47.65 (0.49)女	45.78 (0.37) 1	49.65 (0.23) 1	51.31 (0.07)) 1
联邦新星[47]	21.31 (0.78) 1	28.06 (0.71) 1	45.83 (0.49) 1	55.09 (0.50) 1	56.79 (0.30) 1
脚手架[20]	15.80 (0.84) 1	23.12(0.74) 1	41.29 (0.51) 1	10.00（*失败*）1	10.00（*失败*）1
FedNTD（我们的）	**45.80** (0.53)楼	**54.42** (0.42)女	**58.57** (0.33)女	**60.88** (0.19)女	**62.48** (0.06)女

G.3 ResNet-10 模型的结果

我们报告了一个关于流行架构 ResNet-10 的附加实验。ResNet-10 中的参数数量比主要实验的 2-conv + 2-fc 模型大大约 10 倍。

	美联储平均	脚手架	月亮	FedNTD（我们的）
碎片（s = 2）	36.01	44.59	35.21	**46.27**
碎片（s = 5）	39.21	65.08	51.02	**65.92**
低密度脂蛋白 (a = 0.1)	33.35	38.78	33.57	**49.85**

H 与 KD 的比较

我们通过观察下面损失函数的性能来分析 FedNTD 相对于 KD 的优势。请注意，随着 A 的增加，L(A) 从 $\mathbf{L}_{\text{KD}}$ 移动到 $\mathbf{L}_{\text{NTD}}$，并在 A = 0 和 L_{NTD} 以及 A = 1 时 收缩到 $\mathbf{L}_{\text{KD}}$。

$$\text{A-}_{\text{.N-}} = \mathrm{L}_{\text{CE}}\,(q,\ 1_{\text{y}}) + \mathrm{L(A)}, \tag{17}$$

$$^{\text{L(A)}} = {}^{(1}\_\_{}^{\text{A)}} \bullet \mathbf{L}_{\text{KD}}\,\text{pq}_{\text{l}};\ \text{q}_{\text{g}}\,\text{q} + {}^{\text{A}} \bullet \mathbf{L}_{\text{NTD}}\,\text{pq}_{\text{l}};\ \sim_{g}\ \text{q-} \tag{18}$$

结果在表 12 中显示，达到 L(A) 到 L_{NTD} 会显着提高性能。这种改进支持将非真实类和真实类解耦的效果：使用非真实类信号保存外地分布知识，并从本地数据集中获取关于真实类的新知识。

表 12：通过改变 A 的 CIFAR-10 测试精度。

分区方法 KD „ , KD 7„NTD	„ „ NTD 0.1	0.3	0.5	0.7	0.9
分片 (s = 2)	46.2	46.4	46.9	47.6	48.8	50.2 **52.6**
LDA (a = 0.1)	50.8	50.9	51.4	51.9	52.6	53.6 **54.9**

为了进一步分析 NTD 的影响，我们在通信轮次进行时测量本地模型的性能。结果绘制在图 11 中。请注意，*个性化*性能是在具有相同本地客户端标签分布的测试样本上评估的。

结果表明，虽然 KD 显着提高了服务器模型性能，但 KD 学习的本地模型显示出低得多的本地性能。另一方面，FedNTD 显示出更高的局部性能，这意味着 NTD 成功地解决了蒸馏问题，不会阻碍局部学习。

我们坚持认为，如第 3 节中所建议的，通过在蒸馏损失中丢弃真实类 logits 的这种显着改进来自从本地数据获取新知识和在全局模型中保留旧知识之间更好的权衡。

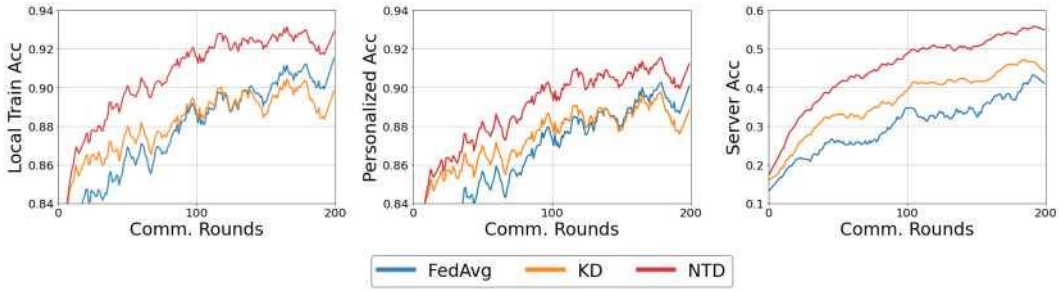


图 11：来自 KD 和 NTD 的 CIFAR-10（分片：*s = 2*）*性能*

I FL 方法的个性化表现

在这里，我们调查 FedNTD 的个性化表现。正如附录 H 中所建议的， 尽管 FedNTD 旨在提高全球收敛性，但它也提高了个性化性能。然而，正如 SCAFFOLD（*绿线*）的学习曲线所示，较低的本地学习性能并不总是导致较差的服务器模型性能。在所有情况下，SCAFFOLD 在每一轮（*第 1 行和第 2 行*）都显示出显着较低的局部性能，但它显着改善了全局收敛（*第 3 行*）。

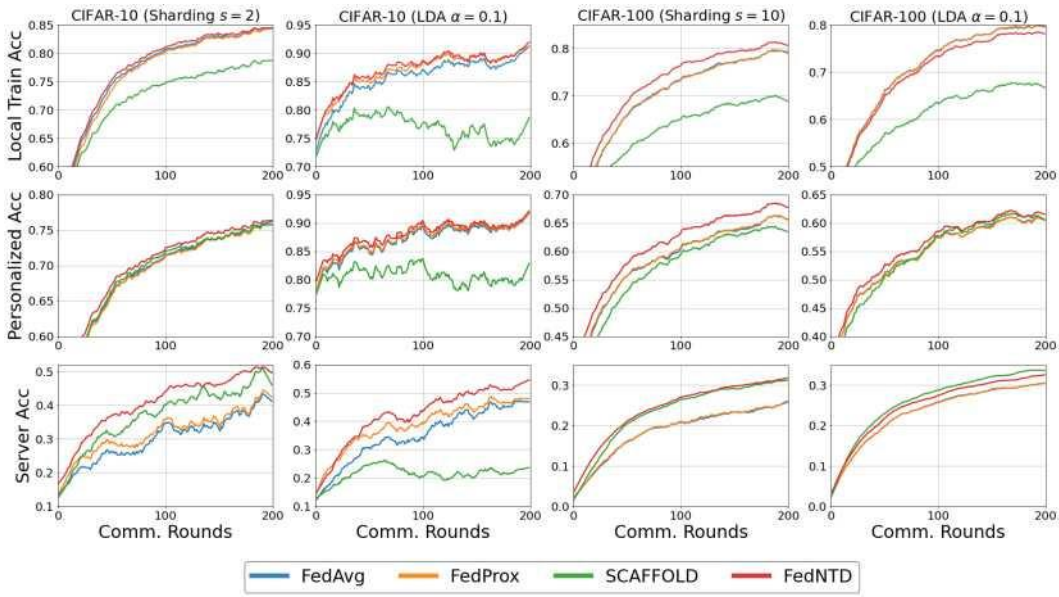


图 12：FL 方法的局部和全局学习曲线。本地模型的准确性在以下方面进行评估：（*本地火车*）-本地私人火车样本，以及（*个性化*）：来自与本地客户端相同标签分布的测试样本

J 与 FedAlign 的比较

在这里，我们将我们的 FedNTD 与最近提出的方法 FedAlign [38]进行比较，该方法与我们的工作动机相同，即局部学习是 FL 性能的瓶颈。在 FedAlign 中，在局部学习目标中引入了校正项以获得泛化良好的局部模型。我们在正式发布的 FedAlign 代码 [2] 上实现了我们的 FedNTD，并使用了[38]中指定的超参数。结果在表 13 中，图 13 显示了它们相应的学习曲线。在我们的实验中，尽管 FedAlign 在某些设置下提高了性能（LDA $a = 0.5$），当异质性水平变得严重时，它的学习就会受到影响。另一方面，即使在这种情况下，FedNTD 也会不断提高性能。

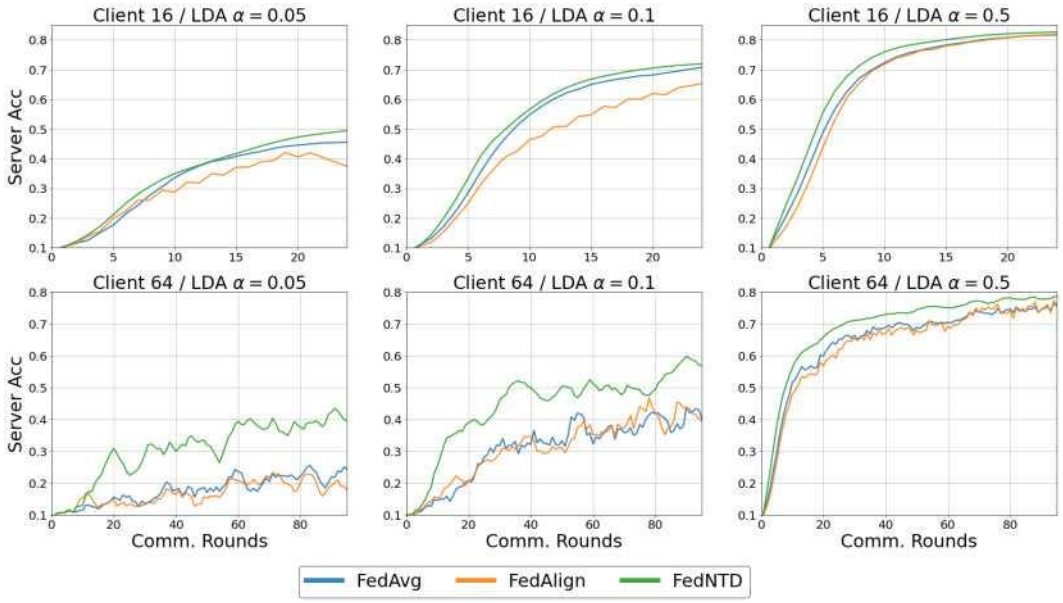


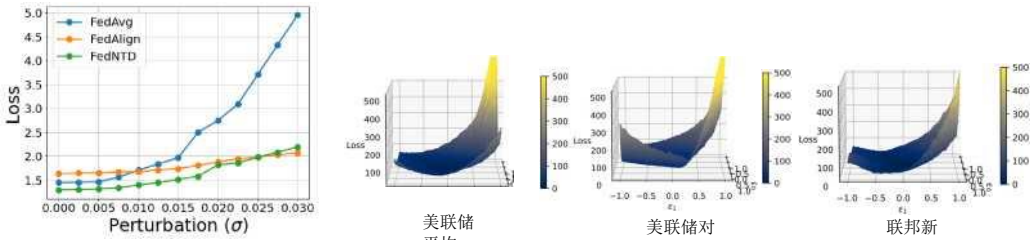
图 13：对应表 13 的学习曲线。

表 13：CIFAR-10 测试精度。每轮沟通有 16 个客户参与。所有实验的局部时期数为 20。

客户端编号 (LDA a)	联邦平均[37]	美联储对齐[38]	FedNTD（我们的）
客户 16 (a = 0.05)	0.4556	0.3743	**0.4943**
客户 16（a = 0.1）	0.7083	0.6532	**0.7195**
客户 16（a = 0.5）	0.8163	0.8185	**0.8266**
客户 64 (a = 0.05)	0.2535	0.1854	**0.3927**
客户 64 (a = 0.1)	0.4247	0.3931	**0.5634**
客户 64 (a = 0.5)	0.7568	0.7698	**0.7846**

FedAlign 引入的损失项旨在在局部训练期间寻找全局分布的分布外普遍性，从而导致跨域的平滑损失景观（= FL 上下文中的异构局部分布）。在图 14 中，我们使用带有高斯噪声的参数扰动和使用 top-2 特征向量轴的可视化来分析损失情况，如[38]中所示。有趣的是，我们的 FedNTD 还平滑了局部景观，这意味着局部训练不需要显着的参数变化来适应其局部分布。我们希望人们能够深入了解损失空间几何中有趣的特性，以解决未来工作中的数据异质性问题。

[2] https://github.com/mmendiet/FedAlign



(a) 扰动引起的损失变化　　(b) Visualized loss landscape with Hessian eigenvectors θ_Q and e_i.

图 14：学习模型的损失空间（Client 16 / LDA a = 0.5）。

FedNTD 超参数的 K 效应

在图 15 中，我们绘制了 FedNTD 超参数对性能的影响。结果表明，尽管 FedNTD 对 *fi 的选择不太敏感，但 r* 太小会显着降低准确性，这可能是由于非真实概率目标过于僵硬。两个超参数对遗忘测度 *F* 的影响如图 16 所示。

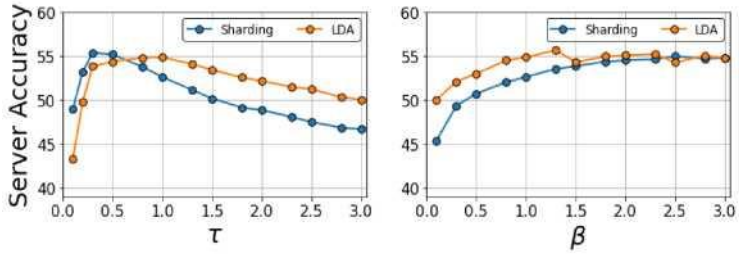


图 15：CIFAR-10（分片：s = 2，LDA：a = 0.1）通过改变 FedNTD 超参数 r 和 ^ 值来测试准确性。

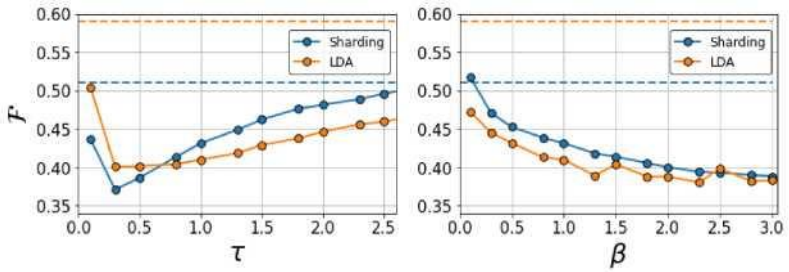


图 16： 通过改变超参数在 CIFAR-10 上忘记 FedNTD 的 *F*。虚线代表基线 FedAvg。

非真实蒸馏的 L MSE 损失

我们探讨了非真实蒸馏的 MSE 损失如何起作用。在表 14 中， MSE 版本 FedNTD (FedNTD (MSE)) 显示出更好的准确性，并且随着 *fi 的*增长，遗忘更少，但在某种程度上，模型发散了；因此无法达到原始的 FedNTD，它利用 softmax 和 KL-Divergence 损失来提取全局模型中的知识。我们将其解释为使用 MSE logits 匹配所有不正确的 logits 过于严格而无法学习全局知识，因为暗知识主要包含在 top-k logits 中。FedNTD 通过使用温度软化的 softmax 来控制类信号。

表 14：CIFAR-10（分片 s=2）通过改变 FedNTD（MSE）和 Fed NTD 的 fi 结果

方法 FedAvg	**FedNTD (MSE) FedNTD**							
£	0.0	0.001	0.005	0.01	0.05	0.1	0.3	1.0
准确性	40.14	40.53	42.39	43.02	44.41	44.27	*失败*	**52.61**
忘记 F	0.59	0.58	0.56	0.55	0.53	0.53	*失败*	**0.43**

M 特征对齐的可视化

为了分析特征对齐，我们将神经元视为基本特征单元，并识别单个神经元的类偏好如下：

$$H = [h_1, h_2 ... h_C], \text{ 其中 } h_c = \wedge O(x_{c,i})\bullet \quad (19)$$

这里，$\mathbf{O}(x_{c,i})$表示神经元对 $_c$类数据 *xi* 的激活，*Nc 是* $_c$类的样本数。对于每个神经元，我们获得最大的类索引 $\text{argmax}_i(\mathbf{H}_i)$，以识别最主要的编码类语义。文献[53]中采用了类似的措施。在图 17 中，我们将最后一层神经元的类别偏好可视化。在 IID 和 NIID（分片 $s = 2$）的情况下，FedNTD 中的特征更加一致。

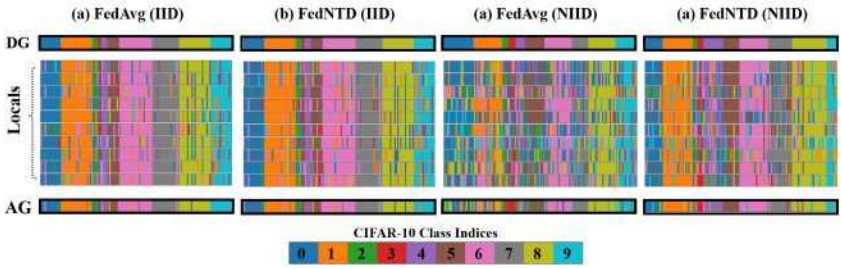


图 17：FedAvg 和 FedNTD 在 CIFAR-10（Sharding=2）上的可视化服务器测试精度。

N 局部特征可视化：Hypersphere

为了弄清楚全局模型中的知识遗忘，我们现在分析局部训练期间全局分布的表示如何变化。为此，我们设计了一个简单的实验来显示单位超球面上的特征变化。更具体地说，我们修改了网络架构以将 CIFAR-10（分片 s = 2）输入数据映射到二维向量并将它们归一化以在单位超球面 $S_1 = \{x \in R^2 : ||x||$上对齐 $_2 = 1\}$。然后我们估计它们的概率密度函数。*全局模型在同质* 本地（iid 分布式）上学习 100 轮通信并分发到*异质* 具有不同本地分布的本地人。结果如图 18

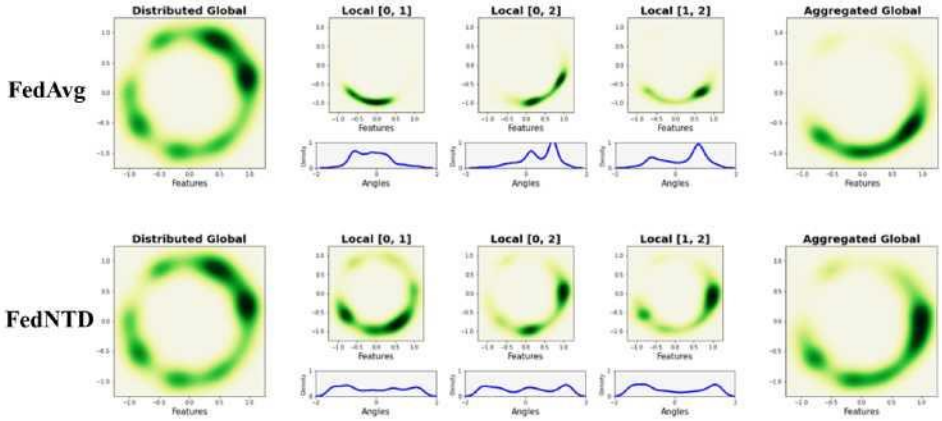
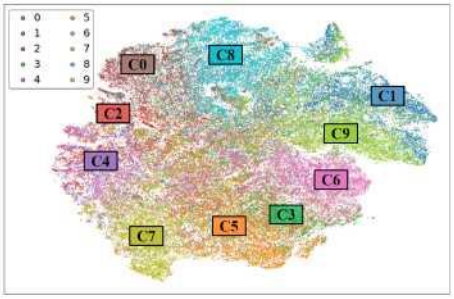


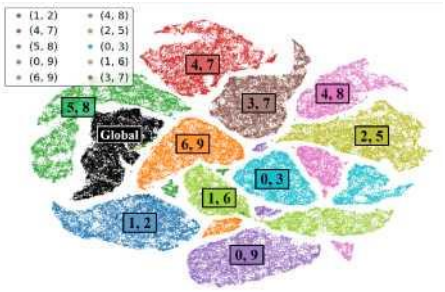
图 18：CIFAR-10（分片 s = 2）测试样本在 S^2上的特征。我们用 R^2中的高斯核密度估计 (KDE)和每个点 $(x, y) \in S^1$的 arctan(y, x)绘制特征分布。分布式全局模型（第一列）在异构局部（中间 3 列）上进行训练，并通过参数平均（最后一列）进行聚合。

O 局部特征可视化：T-SNE

我们进一步对训练有素的局部模型的特征进行了额外的实验。*我们在异构* (NIID) 本地上训练了 100 轮通信的全局服务器模型，并分布在 10 个*同质*(IID) 本地和 10 个*异构* (NIID) 本地上。在同类局部情况下（图 19a 、图 20a ），特征按类聚类，而不管它们是从哪个局部学习的。另一方面，在异构局部情况下（图 19b 、图 20b ），特征被聚类，局部分布被学习。在图 21 中， 我们可视化 FedNTD 对局部特征的影响。

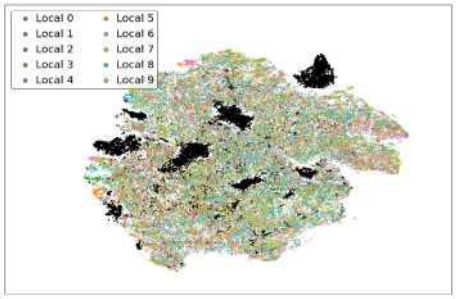


(a) 同质局部变量（按*类别*着色）

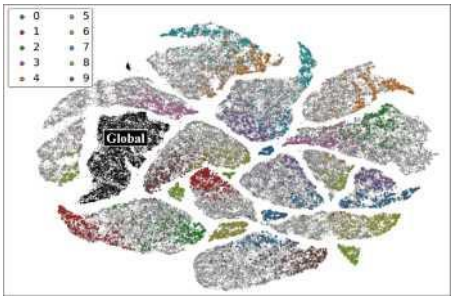


(b) Heterogeneous Locals（由*当地人*着色）

图 19：在对 (a)*同质*局部分布和 (b)*异质*局部分布进行局部训练后，CIFAR-10 测试样本特征的 T-SNE 可视化 。T-SNE 是针对全局模型和 10 个局部模型的测试样本特征一起进行的。

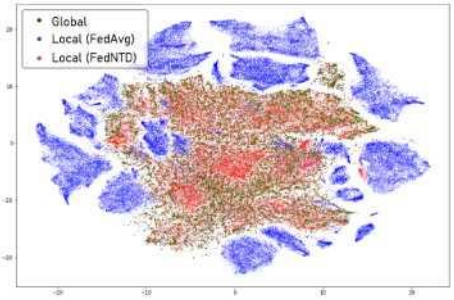


(a) 同质局部变量（按*类别*着色）

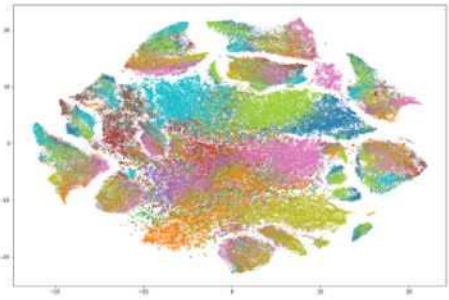


(b) Heterogeneous Locals（由*当地人*着色）

图 20：在对 (a) 均匀局部分布和 (b) 异构局部分布进行局部训练后，CIFAR-10 测试样本上特征区域移动的 T-SNE 可视化。T-SNE 是针对全局模型和 10 个局部模型的测试样本特征一起进行的。



(a) 由不同的算法着色



(b) *按类别着色*

图 21：经过 **FedAvg** 和 **FedNTD 对***异构* 局部分布进行局部训练后，CIFAR-10 测试集样本上的 T-SNE 。T-SNE 是针对全球和 20 个本地模型（10 个 FedAvg 和 10 个 FedNTD）的测试样本特征一起进行的。

P 命题证明 1

证明。由于类梯度 g_i 是相互正交的并且具有统一的权重，从 2-范数的单一不变性我们有：

A(P) = K \ .' \|p k + Pp k w 2 (•) i f = i C = i (p fc (c) + PP " (c)) 2.
w K = i p k + PP " w 2 K C = i (K = i p k (c) + PP " (c)) 2 (20)

= K t £ (p k (c) + PP " (c)) 2 (21)

1 C K 1 C 2 K
- 777 2 K S CE «[c] [(p k + Pp k) 2 1 - (17^) 2 K S EW) [(p k + Pp k) 2 1 (22)

1 c 2 K, •
- K S(Var c e[c][p k + Pp k 1 +(1 + P)j(23)

i C 2 K (r,. i - « k i\ c 2
- (1 + P) 2 K3 S
[jaMC] p + P C|J + K 2 (24)
- (1 +P) 2 K3 S((1 — Ci iV; "'-' p 0 + K
2 ' (25)

其中 (•) 从 p k — V C = 1 p k (c) e k (e A c) 得出，并且 (|) 成立，因为我们假设全球数据分布均匀。也就是说，根据类的对称性，我们有以下等式。

(c)- £ p k (c)- K。 VC k = 1 k = 1 (26)

对方程(25)微分，我们有：

BA(P) (C 2 K k £ (C -(1 + P)) 2
df K 3 (C - I)'■ -" c e[C] [p 1 BP (1 + P) 2 (27)

(C 2 K C 2 C c
K 3 (c - 1) 2 J L Va M c] [p J |^ . 2 (i + p) 3 + 2 (i + f) 2 (28)

(2 C 3 K r k f C 1 \
K 3 (C - 1) 2 J L Var c p[C] [p J (^(i + f) 3 - (i + f) 2) (29)

通过在第一个括号中定义 $M_{\kappa,c,p} > 0$，我们有：

BA C 1
BP - - K,c,H (1 + P)3 (1 + P)2 (30)

对于所有 P ￥ 0。如果 P < **C** /2 — 1，我们有 **C** /(1 + P) ￥ 2，并得到以下所需的不等式：

巴 1
BP^ K, c ,p (1+P) 2。 (31)

Q 命题证明 2

证明。首先，我们展示第一个等式。真实类的总和是：

L_{KL}^{true} ^(yi)qgi(yi) (32)

请注意 $\mathbf{VN}\,\mathrm{L}_{1} = \mathbf{VC} = 1\,\mathrm{S}_{i\,\mathrm{p}\,Sc}$ 和 $i^{\,\mathrm{e}}\,S_{c}\wedge V_{i} = \mathrm{c}$。通过使用这些，我们得到：

$$\mathrm{LKL} = -\,77\,\mathrm{E}\; q_{\mathrm{g}\text{-}i}(v_{\mathrm{i}}) \log \frac{q_{\mathrm{T}^{\text{-}i}}(V_{\mathrm{i}})}{q_{\mathrm{T}^{\text{-}i}}(V_{\mathrm{i}})} \quad \frac{1\ ?}{\text{-NEE}\ q_{\text{-}^{\cdot ic!lo}}\mathrm{g}_{c=1\ i\mathrm{e}S_{c}}} \quad \frac{q^{\wedge}(c)}{q_{\mathrm{T}\text{-}i}(c)} \tag{33}$$

$$= -\mathbf{S}^{C}_{c=1}\mathbf{pels}_{iPS_{c}} \quad \mathbf{S}\; q^{9}{}_{\tau}{}^{\text{-}i(C)} \log \mathbf{IS}_{\mathrm{cl}} \quad \frac{q_{\mathrm{T}^{\text{-}i}}(c)}{q_{\mathrm{T}\text{-}i}(c)} \tag{34}$$

$$= -\,\mathrm{E}\,\underset{c=1}{\overset{C}{\boldsymbol{P}_{c}}}\,{}^{\cdot}\,{}^{\mathrm{E}}{}_{i}\mathrm{PS}_{c} \quad q_{\mathrm{T}\text{-}i}(c) \log \frac{q_{\mathrm{T}^{\text{-}}}}{{}^{i}(c)q_{\mathrm{T}^{\text{-}}}} \tag{35}$$

$$i(c)$$

Next, we derive the not-true part of the Kullback-Leibler divergence:

$$\_L\ \mathbf{L}^{\text{not-true}} \underset{/\ ,\ /\ ,}{\overset{N\ C}{—\ \mathrm{V}\ \mathrm{V}}}\ \text{«}\mathfrak{S}^{\text{-}i}(\mathrm{r}\ 1)\ \mathrm{l}^{\mathrm{N}}\ \mathbf{L}_{\mathrm{KL}} = \frac{q^{l}{}_{T}{}^{i}(c'')}{q_{T^{c}}(d^{1})} \tag{36}$$

By using the double summation technique (★), we have:

$$\text{-}^{N}\mathbf{L}_{\mathrm{KL}} = \overset{nr\ /\text{'not-true}}{2J\ 2\text{-}I\ 2T}\ \overset{CC}{\underset{\mathrm{V}''\ \wedge\mathrm{a\text{-}i}\ 1}{}}\ q_{T^{(C)}} \log \frac{q_{\mathrm{T}\text{-}^{i}}(c')}{q_{\mathrm{T}^{i}}(c')} \quad \overset{CC}{\mathbf{S}\ \mathbf{S}\ \mathbf{S}}\ q_{\mathrm{T}^{\text{-}i}}(c') \log \frac{q^{l}f(c'')}{qT(d')} \tag{37}$$

$$= \mathbf{S}^{CC}\ \mathbf{S}\ \mathbf{S}\ q_{\mathrm{T}\text{-}\mathrm{V}}) \log \frac{q_{T}(c^{1})}{q_{\mathrm{T}^{i}}(c^{1})} = \overset{CC}{\mathbf{S}\mathbf{S}}\ \mathbf{S}\ q_{\mathrm{T}^{i}}(c) \log_{c=1} \frac{qT(c')q}{T^{\text{-}}(c)} \tag{38}$$

$$= (\mathbf{C}\ \text{-}\ 1)\ \mathbf{S}^{C}_{c=1}\ \frac{\mathrm{Scvc}\ |\mathbf{S}_{c}\ \mathrm{I}}{C\ \text{-}\ 1}\ q_{\ \tau}{}^{\mathrm{T}\text{-}i}\ (\mathrm{cq} \log \left[\frac{q^{l}{}_{\tau}{}^{\cdot i}(c)}{q^{y}{}_{T}{}^{i}(c)}\right]) \tag{39}$$

$$= (\mathbf{C}\ \text{-}\ 1)\ \underset{c=1}{\overset{C}{N\boldsymbol{P}_{c}}}\ \bullet\ \mathrm{E}\mathrm{i}\mathrm{R}\mathrm{S}_{c}\ \ q_{\tau}{}^{\mathrm{T}\text{-}i}\ (\mathrm{cq} \log \frac{q_{\mathrm{T}^{\text{-}}}}{{}^{i}(c)q_{\mathrm{T}^{\text{-}}}} \tag{40}$$

$$_{i}(c)$$

因此，我们得到了我们想要的结果：

$$\frac{\mathbf{L}^{\text{not-true}}_{\mathrm{KL}}}{\mathbf{C}\ \text{-}\ 1} = \underset{c=1}{\overset{C}{\mathbf{E}\ \mathrm{P}\mathbf{c}\,\mathrm{E}}}\mathrm{i}\mathrm{R}\mathrm{S}_{c}\ \ q_{\tau}{}^{\mathrm{T}\text{-}i}\ (\mathrm{cq} \log \frac{q_{\mathrm{T}^{\text{-}i}}(c)}{q_{\mathrm{T}^{\text{-}i}}(c)} \tag{41}$$

R 公式 15 的推导

证明。证明的主要部分是著名的光滑函数不等式，它源自泰勒近似。由于 $\mathbf{L}_{\mathrm{i}}\colon \mathbf{W} \in \mathrm{R}^{\mathrm{n}} \wedge \mathrm{R}$ 是光滑函数，我们有

$$\mathbf{L}_{\mathrm{i}}(\mathrm{w}) = \mathbf{L}_{\ \mathrm{i}}(\mathrm{w}_{\mathrm{i}}) + \mathbf{V}\mathbf{L}_{\mathrm{i}}(\mathrm{w}_{\mathrm{i}}) \bullet (w - \mathrm{w}_{\mathrm{i}}) + \int_{0}^{1} (1 - t)(w - \mathrm{W}_{\mathrm{i}})\,\mathrm{J} \bullet \mathbf{V}\,{}_{2}\mathbf{L}_{\mathrm{i}}(\mathrm{w}_{\mathrm{i}} + t(w - \mathrm{w}_{\mathrm{i}}))\,dt \bullet (w - w_{i}) \tag{42}$$

$$= \mathbf{L}_{\mathrm{i}}(\mathrm{w}_{\mathrm{i}}) + \int_{0}^{1} (1 - \mathrm{t})(\mathrm{w} - \mathrm{w}_{\mathrm{i}})\,\mathrm{J} \bullet \mathbf{V}\,{}_{2}\mathbf{L}_{\mathrm{i}}(\mathrm{w}_{\mathrm{i}} + t(w - \mathrm{w}_{\mathrm{i}})) \bullet (w - w_{i})\,dt \tag{43}$$

$$< \mathbf{L}_{\mathrm{i}}(\mathrm{w}_{\mathrm{i}}) + \mathrm{A}\int_{0}^{1} (1 - \mathrm{t})(\mathrm{w} - w_{i})^{J} \bullet (\mathrm{w} - w_{i})\,dt \quad (\mathbf{V}\,{}_{2}\mathbf{L}_{\mathrm{i}}(\mathrm{w}) < \mathrm{A}) \quad \text{1 个}$$

$$= \mathbf{L}_{\mathrm{i}}(\mathrm{wi\ i}) + {}^{A}\|w - w_{i}\|_{2}. \tag{44}$$

□

命题 3 的证明

证明。为了证明这个推论，足以证明下面的极小极大问题是在均匀分布上得到的。

$$\inf_{P \in A_c P \in n} \sup E_{p \sim P}[\| p' - p\|]。 \tag{45}$$

让我们定义 $p \wedge \sup_{P \in n} E_{p \sim P}[\| p' - p\|]$ 为 $F(p)$。首先，我们检查 F 的连续性。那是：

$$|F(p_{2q}) - F(p_1)| \quad \text{支持 } E_{p \sim P}[\| p' - P2 \|] - \sup E_{p \sim P}[\| p' - P 1 1] P \in n \quad P \in n \tag{46}$$

$$< \sup_{\text{笔}} E_{p' \wedge p}[|\, p' - p_2| - \|p' - p_i|\,] < \sup_{\text{笔}} E_{p' \wedge p}[|\, p' - p_2| - \|p' - p_i| \tag{47}$$

$$< \sup_{\text{笔}} E_{p' \wedge p}[|\, p_i - p_{21}|] < \| p_i - p_2 \| 。 \tag{48}$$

因此，由于函数 F 是 1-Lipschitz，它显然是连续的。现在，由于 A_c 是紧致的，我们有一个最小化器 $p_o \in A_c$ 高于极小最大值。因为范数和期望是凸函数，所以 F 是凸函数。因此，对于任意最小值 p_o 和循环 $a = (12 \cdots C) \in S_C$，我们有：

$$F(\text{unif. dist}) = F\left(\frac{1}{\text{我}} \sum a_i(p_o)\right) < \frac{1}{\text{我}} \sum F(a_i(p_o))。 \tag{49}$$

现在，我们论证 $F(a_i(p_o)) = F(p_o)$。根据 F 的定义，

$$F(a_i(p_o)) = \sup_{\text{笔}} E_{p \sim p}[\| p' - a_i(p_o)|\,] = \sup_{\text{笔}} E_{p \sim p}[\,|\, a_i(a_{-i}(p')) - a_i(p_o)\|] \tag{50}$$

$$= \sup_{\text{笔}} E_{p' \wedge p}[\,|\, a_i(a_i(p_1)) - a_i(p_o)\|] \in n \tag{51}$$

$$= \sup_{\text{笔}} \|a_i(a \wedge_i(p')) - p_o\| \, d P(p') \tag{52}$$

$$= \text{支持}[\,|\, a -_i(p') - p_o|\, d P(a_i(a_i(p_1))) \in n A_c \tag{53}$$

$$= \sup_{\text{笔}} \|p'' - p_o\| \, d P(a_i(p_2)) \tag{54}$$

$$= \sup_{\text{笔}} \|p'' - p_o\| \, d P(p_2) = F(p_o)。 \quad (n \text{ 是 } S_C\text{-不变的 } (a_i \in S_C)) \tag{55}$$

从等式等式(49)，我们有：

$$F(\text{unif.dist}) < \frac{1}{\text{二}} \bigwedge F(a_i(p_o)) = \frac{1}{\text{二}} \bigwedge F(p_o) = F(p_o)。 \tag{56}$$

由于 p_o 是最小值，我们可以证明均匀分布也达到了最小值。□.