读Does one-shot give the best shot? Mitigating Model Inconsistency in One-shot Federated Learning.

这篇论文由来自匿名单位的匿名作者撰写，应该还没有正式发表。

引言

所以我们先来了解一下什么是One-shot Federated Learning。通过互联网和LLM，我找到了这个比较好理解的诠释：

标准联邦学习 (FL):

- 多轮迭代: 服务器和客户端之间需要进行很多轮的通信。
- 过程:
  - 服务器将当前的全局模型发送给选定的客户端。
  - 客户端在本地用自己的数据训练这个模型（通常只训练少量几步）。
  - 客户端将模型的更新信息（例如梯度或模型权重变化）发送回服务器。
  - 服务器聚合所有客户端的更新，形成一个新的、改进的全局模型。
  - 重复以上步骤很多次，直到模型收敛。
- 优点: 可以逐步学习，模型精度通常较高。
- 缺点: 通信开销巨大，因为需要反复传输模型或更新。

一次性联邦学习 (OFL):

- 一轮通信: 服务器和客户端之间只有一次主要的通信交互（客户端上传模型）。
- 过程:
  - 客户端（可能从一个初始模型开始，也可能独立从头开始）在本地独立地、长时间地训练自己的模型，直到模型在本地数据上表现良好或者达到预设的训练目标。
  - 训练完成后，每个客户端将自己最终训练好的本地模型一次性上传给服务器。
  - 服务器收集所有客户端上传的最终本地模型。
  - 服务器使用某种聚合策略将这些本地模型一次性合并成一个最终的全局模型。
  - 之后不再有客户端和服务器之间的模型训练相关的通信。
- 优点:
  - 显著降低通信成本: 这是OFL最主要的动机和优势。因为它避免了多轮迭代的巨大通信开销。
  - 简化流程: 协调一次通信比协调多轮要简单。
  - 适用于特定场景: 比如客户端只愿意或只能参与一次模型贡献的场景。
- 缺点:
  - 模型不一致性: 由于客户端长时间独立训练，且它们的数据可能非常不同，导致最终上传的本地模型之间差异巨大，甚至相互冲突。
  - 聚合困难: 简单地平均这些差异巨大的模型，往往得不到好的全局模型。
  - 精度挑战: 传统上，OFL方法的精度通常低于标准的多轮FL，因为聚合步骤面临巨大挑战。

动机

我们回到论文。文中指出现有OFL陷入了一个garbage (inconsistent one-shot local models) in and garbage (degraded global model) out的陷阱，因为：

1. 模型内部不一致性 (Intra-model inconsistency)
   一个一次性局部模型被证明对于具有相同语义的样本会给出不同的预测。
2. 模型间不一致性 (Inter-model inconsistency)。
   来自不同客户端的不同一次性局部模型表现出不同的参数，导致即使对于相同的样本也会产生不同的预测。

说实话我觉得垃圾进垃圾出在这里被滥用了.....这句话应该是描述低价值数据的，而非不一致的模型。

接下来通过数理演算证明了：

1. 一次性局部模型在原始样本 (x, y) 和增强样本 (A(x), y) 上的模型内部不一致性可以表示为：
   $||\Delta_intra||^2 \geq ||(p \cdot \nabla g_a \cdot \nabla A)^T(x - A(x))||^2 > 0$
   其中 $p = \Sigma_{c=1}^{C}(z_c - y_c)$，z 是模型 w_i 对增强样本 A(x) 使用softmax函数激活后的预测，∇g_a 是局部模型 w_i 的梯度，∇A 是数据增强函数的梯度。

   也就是说一次性局部模型中的模型内部不一致性是不可避免的。具体来说，它主要由三个因素引起：(1) 在增强样本上的性能，由 p∇g_a 表示。(2) 数据增强函数的变换属性，由 ∇A 表示。(3) 原始样本和增强样本之间的差异，由 (x - A(x)) 表示。所有这些因素都大于零，导致了模型内部不一致性的存在。
2. 对于任意两个客户端 u 和 v，如果它们拥有相同数量的样本 $n_u = n_v$，那么它们之间模型的一步偏差 $\Delta inter = \nabla W_u - \nabla W_v$ 可以表示为：
   $||\Delta inter||^2 = ||(\eta/N_u)[(n_{u,c}(1 - Z_{u,c})X_{u,c} - n_{v,c}(1 - Z_{v,c})X_{v,c}) - (\Sigma_{c' \in [C_u]} n_{u,c'} Z_{u,c'} X_{u,c'} - \Sigma_{c' \in [C_v]} n_{v,c'} Z_{v,c'} X_{v,c'})]||^2 > 0$
   其中 η 是学习率，$n_{u,c}$ 和 $n_{v,c}$ 是第 c 类别的样本数量，c' 是除了 c 之外的负类。

   那么是说对于任何两个客户端，每个局部训练步骤都会导致一次性局部模型的不一致性。随着局部训练步骤的增加，模型间的不一致性变得更加显著

现有的OFL方法侧重于设计服务器端的聚合机制：

1. 基于优化的方法 (Optimization-based methods)
2. 基于蒸馏的方法 (Distillation-based methods)
3. 生成方法 (Generative methods)
4. 基于选择性集成的方法 (Selective ensemble-based methods)

然而，当前的OFL局部训练策略不可避免地导致“垃圾”一次性局部模型，这给服务器端的聚合带来了挑战。

因此，论文提出了一个新的OFL框架，名为FAFI。它包含两个组件：自对齐局部学习(Self-Alignment Local Learning)和信息感知特征融合推理 (Informative Feature Fused Inference)：

- 在客户端，训练特征提取器，使其能够学习到可以泛化到不同增强样本的不变特征。作者还设计了类别原型学习以获得独特的原型，替换了原始的分类器，从而减轻了预测的负面影响。
- 在服务器端，将所有客户端的原型聚合成一个全局原型。在推理阶段，框架信息感知地融合由局部模型提取的特征，以缓解模型间的不一致性。

方法

自对齐局部学习 (Self-Alignment Local Learning)

导致模型内部不一致性的关键因素是模型无法处理具有相同语义的增强样本，即当 $\nabla g_a > 0$ 时。然而，当前的监督学习范式只能基于标签和原始输入学习固定的语义，缺乏对多样化增强样本的泛化能力。如果模型在任何增强样本上都能表现良好，即 $\nabla g_a = 0$，那么模型内部不一致性将得到缓解。似乎直接采用数据增强可以解决这个问题，然而，受监督训练范式的约束，只要使用标签进行训练 (p ≠ 0)，模型内部不一致性就将不可避免地发生。为此，作者引入自对齐学习来学习不变特征和一个能够泛化到多样化增强样本的无偏分类器。

1. 特征对齐 (Feature Alignment): 为了使模型能够学习更泛化的特征并减少有偏数据的影响，我们专注于学习与自身相关的不变特征，而不是与标签对齐的特征。我们引入了对比学习方法，目标是最小化具有相同语义的样本表示之间的差异，同时最大化与其他所有语义表示的差异，可以表示为：
   $L_s sl = -(1/N_m)\Sigma i = 1^{N_m} log(s(F_{x_i}, F_{x_i^+})/\Sigma_{j \in N_g(y_i)} s(F_{x_i}, F_{x_j}))$

其中 $s(F_{x_i}, F_{x_i^+}) = exp(cos(F_{x_i}, F_{x_i^+})/\tau)$，τ 表示温度参数，$x_i^+$ 表示与 x_i 具有相同标签的样本集，$N_g(y_i)$ 表示与 y_i 不同的样本索引集，cos 函数是余弦相似度。

2. 类别原型学习 (Category-wise Prototype Learning): 虽然对比学习可以实现不变特征，但分类器通常表现出有偏行为并且对数据异构性敏感，导致预测不一致。受到原型学习在各种异构数据场景中成功的启发，可学习的对比原型取代了原始分类器，从而减轻了数据异构性对分类的负面影响。这里的目标是获得一组具有代表性且高度区分性的原型：

    1. 与特征紧密对齐以保留语义信息。
    2. 保持每个原型之间的类别区分性。
       具体来说，我们定义每个客户端维护一组原型 $P_m = P_{m,1}, P_{m,2}, \ldots, P_{m,C}$，其中 C 是类别数。在局部训练期间，更新原型以最小化由局部模型提取的特征与原型之间的对比损失。第 m 个客户端的对比损失可以表示为：

    $L_proto = -(1/N_m)\Sigma_{i=1}^{N_m} log(exp(F_{x_i}^T P_{m,y_i}/\tau)/\Sigma_{c=1}^{C} exp(F_{x_i}^T P_{m,c}/\tau))$

    其中 F_{x_i} 是局部模型为样本 x_i 提取的特征，P_{m,y_i} 是对应于真实类别 y_i 的原型，τ 是温度参数。
    因此，整体的局部训练目标可以表示为：
    $L_local = L_sl + L_proto$
    自监督局部学习损失的目标与缓解模型间不一致性是一致的。并且其中的两个损失是互补的。对于可学习的类别原型 P_m，原始样本和增强样本的预测可以分别表示为 $F_{x_i}^T P_m$和$F_{A(x_i)}^T P_m$。使用余弦相似度作为距离度量，预测差异，即模型内部不一致性，可以表示为：
    $\Delta_intra = -cos(F_{x_i}^T P_m, F_{A(x_i)}^T P_m)$
    $\approx -cos(F_{x_i}, F_{A(x_i)}) P_m, (对应 L_sl)$
    $-F_{x_i}^T P_m, F_{A(x_i)}^T P_m, (对应 L_proto)$
    上述公式的优化目标均对模型内部不一致性有正面的影响。可见，自对齐局部学习损失对模型内部不一致性有优化意义。

信息感知特征融合推理 (Informative Feature Fused Inference):

前面证明了数据异构性引起的模型间不一致性是不可避免的。由于局部模型之间存在显著差异，通过参数级别的聚合难以重构一个表现良好的全局模型，这是导致性能低下的根本原因。那么很容易想到，可以不聚合模型参数，而是融合由不一致的局部模型提取的特征来整合语义，从而减轻数据异构性的负面影响并增强推理能力。此外，注意到不同模型提取的特征在融合过程中表现出差异，为了解决这个问题，作者设计了一种基于注意力的特征融合机制。

特征融合 (Feature Fusion):
具体来说，令 F_m 为第 m 个客户端的局部模型提取的特征。为了融合来自不同客户端的特征，这里设计了一种机制来信息感知地聚合特征。显然，信息量较少、与噪声相似的特征应该被降权，而信息量较多的特征应该被加权。因此，定义重缩放因子 α_m 为：
$\alpha_m = 1 - cos(F_m, F^{N(\mu,\sigma)})$
其中 $cos(,)$ 是余弦相似度，$F^{N(\mu,\sigma)}$ 是由均值为 μ、标准差为 σ 的高斯分布提取的特征。
接下来，使用加权平均聚合来自不同客户端的特征：
$F_fused = \Sigma_{m=1}^{M}(\alpha_m/\Sigma_{v=1}^{M}\alpha_v)F_m$
其中 M 是客户端的数量。

使用全局原型进行推理 (Inference with Global Prototype):
获得融合特征 F_fused 后，我们使用融合特征与具有区分性的全局原型之间的相似性进行预测。具体来说，令
$P_g = (1/M)\Sigma_{m=1}^{M} P_m$ 为从客户端可学习原型聚合而来的全局原型。预测 ŷ 可以表示为：
$\hat{y} = argmax_{c\in[C]} cos(F_fused, P_{g,c})$
这种方法利用了融合的特征和全局原型的区分能力来进行准确预测。

讨论

将原型从客户端传输到服务器，这不会损害隐私安全，因为原型是类别的统计级别信息，不包含个体样本的隐私。此外，类别原型学习旨在捕获同一类别的共同抽象特征，反而可以减少个体样本的暴露风险。

与类似方法的比较:

- FL中的原型 (Prototypes in FL): 现有的基于原型的方法依赖多轮交互来获得代表性的全局原型，这限制了它们在一次性FL场景中的适用性。相比之下，FAFI仅通过一次性聚合就能实现语义对齐的全局原型。

- LLMs中的模型合并 (Model Merging in LLMs): 一些研究认为OFL中现有的聚合过程类似于模型合并，然而：所有这些方法仍然受到“垃圾进，垃圾出”陷阱的影响，因为它们只关注服务器端的合并，忽略了由不一致的预训练模型引起的负面影响；所有这些方法都缺乏隐私保护属性，因为它们需要源数据或与合并后校准高度相关的额外信息。相比之下，FAFI不需要源数据或任何其他辅助信息。

实验

实验证明FAFI：

- 在所有数据集和所有设置上相比基线实现了显著的性能提升（平均10.86%）
- 在不同的客户端规模下始终保持最佳性能
- 在产生较低通信开销的同时，实现了比所有其他基线更高的性能。
- 更大的批量大小会带来性能提升，但需要更多资源，即使使用最小的批量大小，性能仍然具有竞争力
- 在域漂移和多任务场景中表现出局限性，在这些场景中，静态的类别表示无法适应冲突的特征分布或发散的任务目标，然而该局限性与其他基于原型的方法共享

结论

本文提出了一个新颖的OFL框架，即FAFI，以缓解模型不一致性。本文首先通过经验和理论揭示了现有OFL方法中在模型内部和模型间两个层面上存在的不一致性。为此，作者提出了用于语义对齐特征提取器和独特性原型的自对齐局部训练，并使用信息感知特征融合推理来融合特征以进行推理。在三个数据集上的广泛实验证明了这种方法的优越性。