Fair Resource Allcation in Federated Learning

大纲

- 背景
- 基本思想
- 传统联邦学习
- 算法
  - `q-FFL`
  - `q-FedAvg`
- 一些想法

背景

现有的联邦学习训练方法可能会导致最终训练出来的模型不适合某些设备

- 参与到一个联邦学习任务中的设备一共有1000台，其中有900台是手机，90台是电脑，10台是嵌入式设备(智能手表手环等)，那么根据传统的联邦学习训练的方法（随机选择一些设备来进行训练），那么最终训练出来的模型就很适合手机，但是不适合电脑和嵌入式设备，表现在准确度方面很差。
- 另外一种情况就是不同设备可以用来训练的数据不同。例如从存储空间方面来看，电脑明显比手机和嵌入式设备能容纳更多训练数据；从数据的可获取度来看，嵌入式设备（运动手环）能够收集到的数据也比手机更多。因此数据集大小的不对等，也有可能造成最终训练出来的模型，更适合那些数据集大的设备，而在数据集小的设备上表现很差

基于这种问题，作者提出了两种训练方法：

- `q-FFL`：适用于大规模的联邦学习网络
- `q-FedAvg`：更轻量级的

基本思想

作者的基本思想是，给予经验损失比较大的设备更高的权重

传统联邦学习

先前的联邦学习训练方法相当于最小化以下的损失函数：

$$min_w f(w) = \sum_{k=1}^{m} p_k F_k(w)$$

其中p_k表示第k个设备所具有的权重，所以$\sum_k p_k = 1$，第k个设备的损失函数是$F_k(w) = \frac{1}{n_K}\sum_{j_k=1}^{n_k} l_{j_k}(w)$，这里$n_k$表示在第$k$个设备中可以用来训练的样本，$l_{j_k}(w)$表示在第$k$个设备中第$j_k$个数据在经过神经网络之后得到的一个损失，然后$p_k = \frac{n_k}{n}$，$n = \sum_k n_k$，就是将第$k$个设备上可以用来训练的样本数量与这一轮被选中的所有设备的样本数的总和的比例作为它的权重

然后这种简单的随机选择方法就会产生以上问题

公平性的定义

如何评判两个模型谁更公平

例如用不同的联邦算法训练出两个全局模型w和$\tilde{w}$，联邦网络中有m个节点，如果

$$Var(a_1, a_2, \ldots, a_m) \leq Var(\tilde{a_1}, \tilde{a_2}, \ldots, \tilde{a_m})$$

那么认为w比$\tilde{w}$更公平，其中a_k表示第k个设备的一个性能的度量，可以准确度等方式来表示，$Var()$表示方差

q-FFL

基本思想是，给予那些性能差的设备更高的权重

相当于最小化以下损失函数：

$$min_w f_q(w) = \sum_{k=1}^{m} \frac{p_k}{q+1} F_k^{q+1}(w) = \sum_{k=1}^{m} p_k \frac{F_k^{q+1}(w)}{q+1}$$

这里一个比较关键的超参数是q，当q等于0时，相当于不引入公平性作为考虑的范畴；当$q > 0$时，q值越大，说我们更强调那些性能比较差的设备所造成的影响

具体的算法为：

Algorithm 1 q-FedSGD

Input: $K, T, q, 1/L, w^0, p_k, k = 1, \cdots, m$
for $t = 0, \cdots, T-1$ **do**
  Server selects a subset S_t of K devices at random (each device k is chosen with prob. p_k)
  Server sends w^t to all selected devices
  Each selected device k computes:

$$\Delta_k^t = F_k^q(w^t)\nabla F_k(w^t)$$
$$h_k^t = qF_k^{q-1}(w^t)\|\nabla F_k(w^t)\|^2 + LF_k^q(w^t)$$

  Each selected device k sends Δ_k^t and h_k^t back to the server
  Server updates w^{t+1} as:

$$w^{t+1} = w^t - \frac{\sum_{k\in S_t}\Delta_k^t}{\sum_{k\in S_t} h_k^t}$$

end for

其中Δ_k^t表示第t轮第k个设备上传的更新，h_k^t为对应的步长的倒数，使用这个步长的意义在于，根据更新梯度的斜率，来调整更新步长。因为h_k^t是函数$\frac{f^{q+1}(w)}{q+1}$的梯度的斜率的一个上界（利普希兹常量），当h_k^t很小的时候，代表着目标函数还是有着下降的可能，因此以比较大的步长进行更新；当h_k^t很大的时候，说明梯度逐渐变小，目标函数可能来到了最低点，因此将步长调小一点

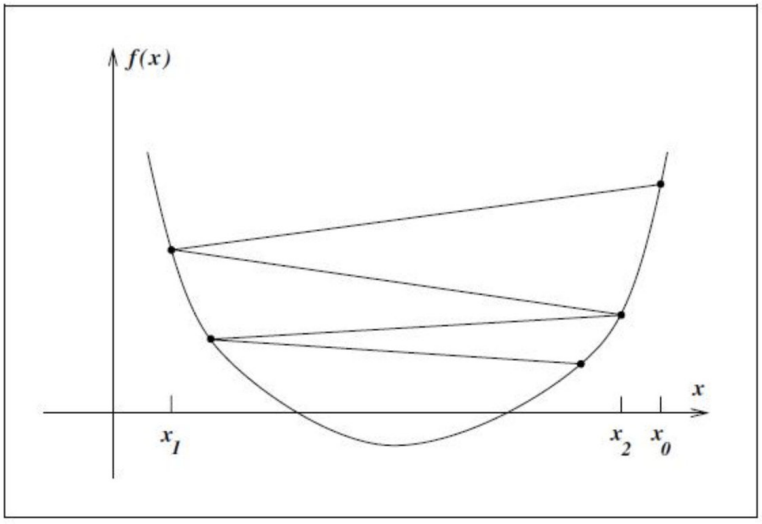

Figure 3.2 Insufficient reduction in f.

q-FedAvg

该方法的与`q-FFL`的区别在于，`q-FFL`在设备上只进行一次神经网络的前向传播和反向传播，然后直接将得到的梯度更新返回，而`q-FedAvg`则是先在设备上进行训练得到一个新的模型，然后将新模型与旧模型的差值作为梯度更新返回，后者与传统的联合平均算法`FedAvg`一致，但是在这个问题上，如果直接用联合平均的算法，那么最后优化的只是原来的那个损失函数，而不是作者修改过的那个损失函数。因此作者提出了一种新的`q-FedAvg`算法

先在设备上使用SGD方法来对模型进行E轮次的训练，得到$\overline{w}^{t+1}$，然后用$L(w^t-\overline{w}^{t+1})$代替$\nabla F_k(w^t)$

Algorithm 2 q-FedAvg

Input: $K, E, T, q, 1/L, \eta, w^0, p_k, k=1,\cdots,m$
for $t=0,\cdots,T-1$ **do**
  Server selects a subset S_t of K devices at random (each device k is chosen with prob. p_k)
  Server sends w^t to all selected devices
  Each selected device k updates w^t for E epochs of SGD on F_k with step-size η to obtain $\bar{w}_k^{t+1}$
  Each selected device k computes:

$$\Delta w_k^t = L(w^t - \bar{w}_k^{t+1})$$
$$\Delta_k^t = F_k^q(w^t)\Delta w_k^t$$
$$h_k^t = qF_k^{q-1}(w^t)\|\Delta w_k^t\|^2 + LF_k^q(w^t)$$

  Each selected device k sends Δ_k^t and h_k^t back to the server
  Server updates w^{t+1} as:

$$w^{t+1} = w^t - \frac{\sum_{k\in S_t}\Delta_k^t}{\sum_{k\in S_t}h_k^t}$$

end for

可以这样理解，`q-FFL`作用的是一次迭代产生的梯度更新，而`q-FedAvg`作用的是多次迭代产生的梯度更新，这两者在本质上都是梯度更新，只不过后者进行了多次的训练，能够保证更高的准确率。这也是为什么第一个方法适用于大规模的网络，因为在大规模的网络中参与联邦学习的设备比较多，在数据集足够的情况下不需要再同一个设备中进行多次迭代的训练

一些想法

基于这种思路的其他的实现方式

在联合的时候重新分配权重

首先虽然该paper在文章的前面说基本思路是：“给予经验损失比较大的设备更高的权重”，但他实际上不是这么做的

可以在联合平均的时候，除了上传梯度更新Δw_k^t，另外上传$L_k^t(w)$，然后联合的时候：

$$w^{t+1} = w^t - \frac{\sum_{k \in S_t} \frac{L_k^t(w)}{\sum_{k \in S_t} L_k^t(w)} \Delta w_k^t}{step}$$

- 或者对经验损失做一个$softmax$，给予经验损失比较大的设备更高的权重
- 又或者是根据在设备上的数据集大小来做一个$softmax$，给予数据集数量比较小的设备更高的权重
- 但是这个想法有一个缺点：上面的联合平均的方法相当将参与联邦学习的设备都复制几份，只是性能好的设备复制的副本少，性能差的设备复制的副本多，这样子虽然同样的副本只需要训练一次，但是还是没解决样本数据少的问题，而这个问题在`q-FFL`算法中也是存在的
- 关键在于如何对损失函数进行量化

换一个目标函数

这个算法的出发点就是在作者将目标函数改为了$L(w) = \frac{f^{q+1}(w)}{q+1}$，但这个目标函数不是唯一的，只要保证$\frac{\partial L(w)}{\partial f(w)} > 1$便可

该算法可应用的领域

- 这个`q-FFL`算法（训练方法）也可以运用到一些经典的机器学习领域，比如说分类问题，如果在一个数据集中某个属性的数据过于少，那么如果使用传统的机器学习算法，最终训练出来的模型也许对于这些数据的性能非常不理想，这样的训练方法对于那些数据来说是不公平的，同样可以运用上面的一个算法，来进行更加公平的训练
- 社交网络

一些问题

- 进行联合平均时的公式：

$$w^{t+1} = w^t - \frac{\sum_{k \in S_t} \Delta_k^t}{\sum_{k \in S_t} h_k^t}$$

  不应该是这样子？：

$$w^{t+1} = w^t - \sum_{k \in S_t} \frac{\Delta_k^t}{h_k^t}$$

- 为什么要用$L(w^t - \overline{w}^{t+1})$来代替$\nabla F_k(w^t)$