УДК 004.7

Ensemble aware optimization in federated learning

Т.Д. Белинский[1], А.Р. Зайнуллин[1], Д.Д. Феоктистов[2, 3]

[1]Московский физико-технический институт (национальный исследовательский университет)
[2]Факультет вычислительной математики и кибернетики (ВМК МГУ)
[3]Yandex Research ML Residency

Пока традиционные методы ансамблирования обучают модели независимо, в данный работе исследуется гибридный подход, который комбинирует независимое обучение с "joint loss"-оптимизацией. Неожиданно эта идея имеет потенциал в нескольких вариантах обучения ансамблей. Первый – это "aligned training", где члены обучаются строить наиболее скоординированные прогнозы. Второе направление относится к федеративному обучению, где мы стремимся обучать ансамбль как глобальную модель. Мы предлагаем новый алгоритм для решения этих задач, через эксперименты показывая, что он превосходит бейзлайны в сценариях федеративного обучения при высокой неоднородности данных.

С точки зрения оптимизации комбинация независимого обучения с "joint loss"-оптимизацией может выглядеть следующим образом:

$$\min_{\theta_1 \dots \theta_N} \mathcal{L}\left(\frac{1}{N}\sum_{i=1}^{N} f(\theta_i)\right) + \mu \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}(f(\theta_i)) \qquad (1)$$

где $f(\cdot)$ – это некоторая модель с обучаемыми параметрами $\theta_1, \dots \theta_N$, а $\mathcal{L}$ – функционал эмпирического риска. Мы предлагаем следующий алгоритм для решения данной задачи оптимизации (1)

Algorithm 1 Smart ensemble

1: **Вход:** Начальные параметры $\theta_1^0, \dots, \theta_N^0$, количество итераций K, количество устройств N
2: **for** $k = 0, \dots, K-1$ **do**
3: **for** $i = 1, \dots, P$ **do**
4: $\theta_m^{k+1} = \theta_m^k - \gamma \nabla_{\theta_m^k}(\mathcal{L}(f_m(\theta_m^k)))$
5: **end for**
6: $\theta_m^{k+1} = \theta_m^k - \gamma \nabla_{\theta_m^k}\left[\mathcal{L}\left(\frac{1}{N}\sum\limits_{m=1}^{N} f_m(\theta_m^k)\right)\right]$
7: **end for**

Данный подход имеет явное преимущество над имеющимися методами в FL, такими как FedAVG [McMahan et al., 2023] и SCAFFOLD [Karimireddy et al., 2020] по стоимости коммуникаций, ведь вместо усреднения всех весов моделей, использует лишь выходы моделей.

В экспериментах мы обучаем N клиентов. Для внесения неоднородности данные распределяются следующим образом. Для каждого клиента выбираются нескольких доминантных классов. Затем датасет делится для каждого клиента на три части: 80% — обучение, 10% — валидация, 10% — коммуникация. В каждой из этих частей выполняется стратифицированная выборка: с прВ каждой из этих частей выполняется стратифицированная выборка: отдающая предпочтение доминантным классам. Таким образом, каждый клиент получает три выборки с одинаковым распределением классов для одного клиента, но различным между клиентами.

Части для коммуникации объединяются в один датасет, в котором классы распределены равномерно за счёт непересекающихся доминантных классов.

Наши эксперименты, проведенные с использованием моделей ResNet18 и SWIN, а так же датасетов CIFAR-10 и ImageNet, показывают что в условиях экстремальной гетерогенности наш метод превосходит по метрике точности имеющиеся бейзлайны.

Результаты одного из экспериментов представлены на рисунке 1. В нем мы решали задачу бинарной классификации используя две модели ResNet18 и выборку изображений из датасета CIFAR-10

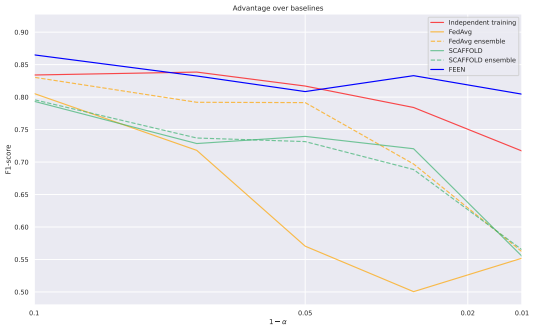

Figure 1: accuracy в зависимости от степени гетерогенности данных. Чем меньше выражение $1 - \alpha$ (правая область графика) тем с более неоднородными данными мы имеем дело

References

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pages 5132–5143. PMLR, 2020.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data, 2023. URL https://arxiv.org/abs/1602.05629.