Ensemble Aware Optimization In Federated Learning

Белинский Тимофей МФТИ

Зайнуллин Амир МФТИ

Феоктистов Дмитрий

ВМК МГУ, Yandex Research ML Residency

18 мая 2025 г.

План

1. Постановка задачи
2. Алгоритм
3. Некоторые результаты
4. Выводы и дальнейшее исследование

Постановка задачи

С точки зрения оптимизации, обучение ансамбля из M моделей с параметрами $x_1, ...x_M$, можно сформулировать как следующую задачу:

$$\min_{x_1,...,x_M \in \mathbb{R}^d} \left\{ L_{ind}(x_1, ..., x_M) := \frac{1}{M} \sum_{m=1}^{M} \sum_{j=1}^{D} \mathcal{L}(f(a_j, x_m), y_j) \right\}, \quad (1)$$

Где (a_j, y_j) это элемент датасета (A, y) размера D, $\mathcal{L}(\cdot, \cdot)$ - функция потерь, $f(a_j, x_m)$ предсказания модели с весами x_m для объекта a_j с меткой y_j.

Постановка задачи

Другой подход - обучать модели таким же образом, каким они будут производить инференс, т.е. усреднять предсказания до подсчета функции потерь. Этот подход известен как joint обучение

$$\min_{x_1,\ldots,x_M \in \mathbb{R}^d} \left\{ L_{joint}(x_1, \ldots, x_M) := \sum_{j=1}^{D} \mathcal{L}\left(\frac{1}{M} \sum_{m=1}^{M} f(a_j, x_m), y_j \right) \right\}, \quad (2)$$

Постановка задачи

Теперь подведем нашу задачу к контексту федеративного обучения (Konečnỳ и др., “Federated optimization: Distributed machine learning for on-device intelligence”).
Каждый агент обучается на своем приватном датасете ((A^m, y^m) размера D_m), а joint компонента считается на другом датасете (A^{joint}, y^{joint}) размера D, например публичный датасет или собранный из кусков, которыми скинулись агенты.

$$\min_{x_1,\dots,x_M \in \mathbb{R}^d} \left\{ \frac{1}{M} \sum_{m=1}^{M} \sum_{j=1}^{D_m} \mathcal{L}(f(a_j^m, x_m), y_j^m) + \lambda \sum_{j=1}^{D} \mathcal{L}\left(\frac{1}{M} \sum_{m=1}^{M} f(a_j^{joint}, x_m), y_j^{joint} \right) \right\}, \quad (3)$$

Алгоритм

Algorithm 1 FEEN

Require: Initial parameters $\theta_1^0, \ldots, \theta_N^0$, number of iterations K, number of devices P

1: **for** $k = 0$ to $K - 1$ **do**
2: **for** $m = 1$ to P **do**
3: $\theta_m^{k+1} \leftarrow \theta_m^k - \gamma \nabla_{\theta_m^k} \left(\mathcal{L}\left(f_m(\theta_m^k) \right) \right)$
4: **end for**
5: $\theta_m^{k+1} \leftarrow \theta_m^k - \gamma \nabla_{\theta_m^k} \left[\mathcal{L}\left(\frac{1}{N} \sum_{m=1}^{N} f_m(\theta_m^k) \right) \right]$
6: **end for**

Мотивация

Многие FL алгоритмы обучения имеют стоимость коммуникации пропорциональную весу модели. Например FedAvg (McMahan и др., *Communication-Efficient Learning of Deep Networks from Decentralized Data*)
SCAFFOLD (Karimireddy и др., *SCAFFOLD: Stochastic Controlled Averaging for Federated Learning*)
В нашем алгоритме стоимость коммуникации в нашем случае пропорциональна размеру shared датасета.

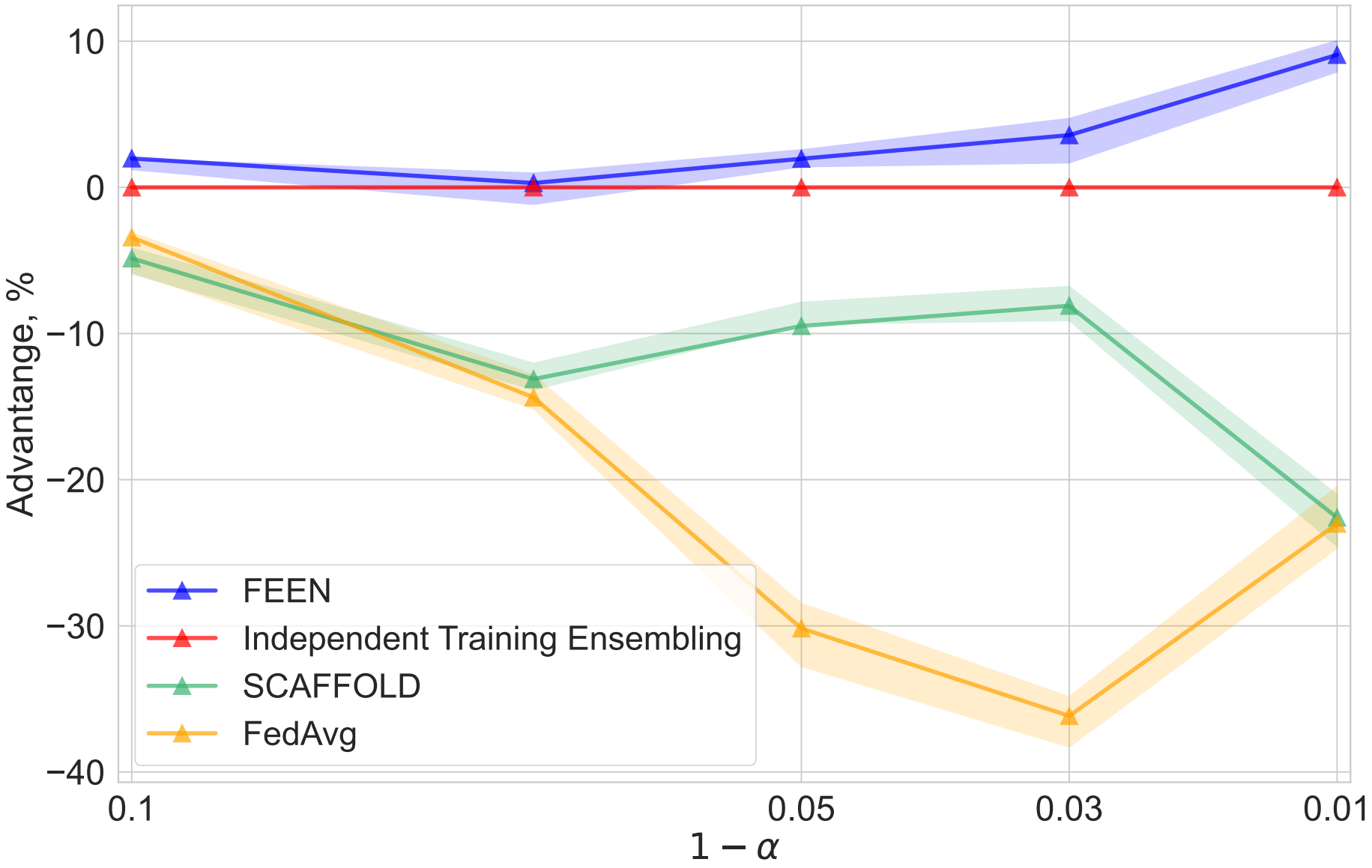
Некоторые результаты

- В условиях гомогенности данных метод дает прирост в качестве, но небольшой
- Результат эксперимента становится слабо заметен, если модель слишком производительная. Для ее ослабления, будем уменьшать размер обучающихся данных.
- В условиях гетерогенности эффект виден более выразительно

Гетерогенность

- Пусть 2 агента решают задачу бинарной классификации
- Пусть у 1-го агента доля объектов перового класса в обучающей выборке α, и, соответственно $1 - \alpha$ – второго
- Если агенты поделятся данными для коммуникационной выборки, то она будет гомогенной
- Тест проводим на гомогенной выборке

Toy example


10
0
−10
−20
−30
−40
Advantange, %
0.1
0.05
0.03
0.01
1 − α
FEEN
Independent Training Ensembling
SCAFFOLD
FedAvg

Выводы

- методы без объединения в ансамбли терпят неудачу из-за крайней неоднородности
- Независимая сборка ансамблей дает значительно лучшие результаты,
- Предлагаемый нами метод FEEN сокращает переобучение локальных моделей с помощью коммуникаций

Дальнейшее исследование

- Подробнее изучить постановку с гетерогенностью (*тюнить параметры коммуникации, реализовать мультиклассовую постановку*)
- Попробовать провести обучение в других условиях (*поменять датасет, модель*)

Список литературы I

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S. J., Stich, S. U., & Suresh, A. T. (2021). SCAFFOLD: Stochastic Controlled Averaging for Federated Learning. https://arxiv.org/abs/1910.06378

Konečnỳ, J., McMahan, H. B., Ramage, D., & Richtárik, P. (2016). Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*.

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2023). Communication-Efficient Learning of Deep Networks from Decentralized Data. https://arxiv.org/abs/1602.05629

Спасибо за внимание!