АССОЦИАЦИЯ
БОЛЬШИХ ДАННЫХ

КОНФИДЕНЦИАЛЬНЫЕ ВЫЧИСЛЕНИЯ И ДОВЕРЕННЫЕ СРЕДЫ ИСПОЛНЕНИЯ

Федеративное обучение (Federated learning)

Москва
2025

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Как известно, совершенствование алгоритмов машинного обучения (англ. Machine Learning, ML) в значительной степени зависит от качества и объема обучающих датасетов. Вместе с тем использование данных, имеющихся в открытом доступе, зачастую оказывается малоэффективным при решении специализированных задач. Преодолеть нехватку информационного «сырья» для обучающих датасетов можно путем их наращивания за счет данных, распространение и использование которых подпадает под законодательные или коммерческие ограничения, — разумеется, при условии надежной защиты этих данных.

Как уже отмечалось в предыдущей статье о технологии SMPC, на текущий момент получили признание несколько стратегий защиты данных на стадии использования. Одной из них является предложенная[1] в 2016 году парадигма машинного обучения, известная сегодня как федеративное обучение (англ. Federated Learning, FL). Она решает проблемы, связанные с невозможностью или нежеланием формирования централизованного датасета несколькими владельцами данных, при этом позволяя им обучать совместную ML-модель. В самых общих чертах FL выглядит следующим образом: каждый участник обучения (FL-клиент) выполняет обучение локальной подмодели на своих вычислительных ресурсах с использованием имеющихся только у него данных (исходные данные при этом никому из других участников не передаются), после чего параметры полученных локальных подмоделей агрегируются. Результатом этой работы становится формирование параметров глобальной модели, концентрирующей в себе знания, извлеченные из совокупности обучающих примеров всех участников. Например, несколько банков могут таким образом совместно обучить модель антифрода, характеристики которой, благодаря большему числу фрод-транзакций в обучающей выборке, будут существенно лучше характеристик локальных моделей, обученных каждым банком в отдельности.

[1] H. B. McMahan, E. Moore, D. Ramage, Federated learning of deep networks using model averaging, arXiv preprint arXiv: 1602.05629, 2016.

КОНФИДЕНЦИАЛЬНЫЕ ВЫЧИСЛЕНИЯ И ДОВЕРЕННЫЕ СРЕДЫ ИСПОЛНЕНИЯ
Федеративное обучение (Federated learning)

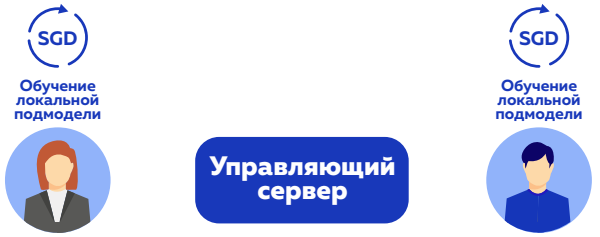
ОБЗОР ТЕХНОЛОГИИ

Ключевой особенностью концепции федеративного обучения и важным шагом к обеспечению конфиденциальности данных является отсутствие передачи исходных данных в ходе совместного обучения ML-модели. Как же осуществляется процесс обучения глобальной модели? Для ответа на этот вопрос следует рассмотреть FL как задачу оптимизации, распределенную по всем FL-клиентам.

Предположим, что у двух владельцев данных — Алисы и Боба — имеются размеченные наборы данных. Тогда процесс федеративного обучения ими совместной модели можно представить в виде следующей последовательности шагов:

Шаг 1. Инициализация: управляющий сервер (в этой роли может выступать некоторый доверенный сервер или один из участников обучения) генерирует начальные значения параметров глобальной модели и рассылает их Алисе и Бобу.

Шаг 2. Локальное обучение: Алиса и Боб, получив от управляющего сервера глобальные параметры, инициализируют ими свои локальные подмодели и проводят раунд обучения. Обычно для этого используются вариации стохастического градиентного спуска (SGD).



Шаг 3. Отправка параметров: после проведения локального обновления Алиса и Боб отправляют свои параметры на управляющий сервер.

Шаг 4. Агрегация: на сервере выполняется агрегирование локальных параметров с использованием заранее выбранного метода (подробнее о них будет рассказано в главе «Структура методов и протоколов»).

Шаг 5. Обновление параметров: в результате агрегации формируются новые параметры глобальной модели, которые отправляются Алисе и Бобу.

Шаг 6. Повторение: действия шагов 2–5 выполняются, пока не будет выполнен критерий остановки процедуры обучения.

Строго говоря, задачу распределенной оптимизации необходимо рассматривать как задачу минимизации глобальной функции потерь на основе данных, хранящихся локально у нескольких участников. В общем случае для K FL-клиентов, у каждого из которых имеется свой набор данных D_k, $k \in \overline{1;K}$, состоящий из n_k, $k \in \overline{1;K}$, пар $(\overline{x}, y)$, где $\overline{x}$ — независимые переменные, y — целевой признак, $n = \sum_{k=1}^{K} n_k$ — совокупное количество пар, необходимо минимизировать глобальную функцию потерь $L(\Theta)$:

$$min_{\Theta \in R^d} L(\Theta) = \frac{1}{n} \sum_{i=1}^{n} l(x_i, y_i, \Theta)$$

Здесь $l(x_i, y_i, \Theta)$ — функция потерь для обучающего экземпляра (x_i, y_i), зависящая от параметра модели Θ. Очевидно, что ни один участник обучения не имеет доступа ко всем данным и не может вычислить глобальную функцию потерь напрямую, поэтому каждый FL-клиент оптимизирует локальный вариант модели на наборе данных D_k, вычисляя локальную функцию потерь:

$$L_k(\Theta) = \frac{1}{n_k} \sum_{(x_i, y_i) \in D_k}^{n} l(x_i, y_i, \Theta)$$

В итоге выражение для вычисления глобальной функции потерь приобретает следующий вид:

$$L(\Theta) = \sum_{i=1}^{K} \frac{n_k}{n} L_k(\Theta)$$

СТРУКТУРА МЕТОДОВ И ПРОТОКОЛОВ

Благодаря появлению концепции федеративного обучения открылась новая обширная область исследований, что привело к выделению FL в самостоятельное направление машинного обучения, включающее в себя различные вариации. Так, по типу распределения данных между участниками взаимодействия и способам формирования глобальной модели принято выделять:

- горизонтальное федеративное обучение (англ. Horizontal Federated Learning, HFL);
- вертикальное федеративное обучение (англ. Vertical Federated Learning, VFL);
- федеративное трансферное обучение (англ. Federated Transfer Learning, FTL).

Горизонтальное федеративное обучение

Горизонтальное федеративное обучение можно применять, когда FL-клиенты имеют одинаковое пространство признаков, но разные образцы данных (сущности), то есть когда между данными клиентов можно провести воображаемую горизонтальную границу (см. рис. 1). Практическую ценность представляют случаи, когда несколько организаций или устройств собирают идентичные признаки, описывающие разные сущности.

Например, перспективным для HFL является применение в медицине, когда на медицинских изображениях, полученных в клиниках по всему миру, обучаются модели, помогающие проводить диагностику заболеваний. Кроме того, HFL может найти применение в банковской сфере, например, для улучшения скоринговых или антифродовых моделей. Примером практического применения данного подхода может служить обучение компанией Google своей модели предсказания следующего слова при использовании клавиатуры GBoard непосредственно на Android-устройствах.

Основным мотивационным фактором для клиентов при организации достаточно сложной процедуры обучения является получение более качественных моделей. Важно при этом учитывать, что, в случае статистической однородности обучающих выборок FL-клиентов в выбранном пространстве признаков, глобальная модель может не продемонстрировать прироста метрик качества. Однако при наличии статистических особенностей локальных датасетов модель, обученная на совокупности данных, гарантированно будет обладать большей обобщающей способностью, что достигается благодаря большему разнообразию обучающих образцов. Вместе с тем такая модель с высокой вероятностью будет иметь меньшую точность внутри специфического домена по сравнению с моделью, обученной локально.

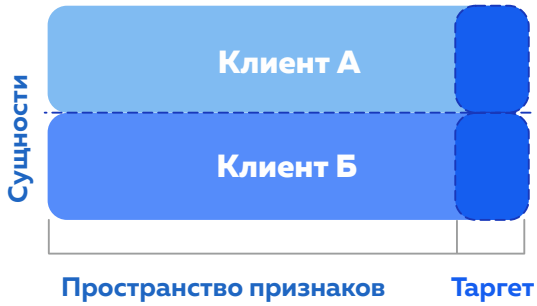


Рис. 1. Схема разделения данных при HFL

HFL является базовым вариантом FL, позволяющим проводить обучение моделей любых архитектур — от линейных моделей до глубоких нейронных сетей.

Выделяют два основных сценария применения HFL:
Кросс-секционное (cross-silo) — в котором подразумевается взаимодействие владельцев данных, заинтересованных в получении более точной модели по сравнению с той, что обучена на их локальных данных.

КОНФИДЕНЦИАЛЬНЫЕ ВЫЧИСЛЕНИЯ И ДОВЕРЕННЫЕ СРЕДЫ ИСПОЛНЕНИЯ
Федеративное обучение (Federated learning)

Кросс-устройственное (cross-device) — когда дообучение глобальной модели происходит на устройствах, непосредственно продуцирующих данные. В этом сценарии заинтересованными являются две стороны: вендор программного обеспечения, содержащего исходную модель, поскольку он обновляет интегрированную модель на основе новых данных, и владелец устройства, который, благодаря адаптации модели, получает более точные результаты ее работы.

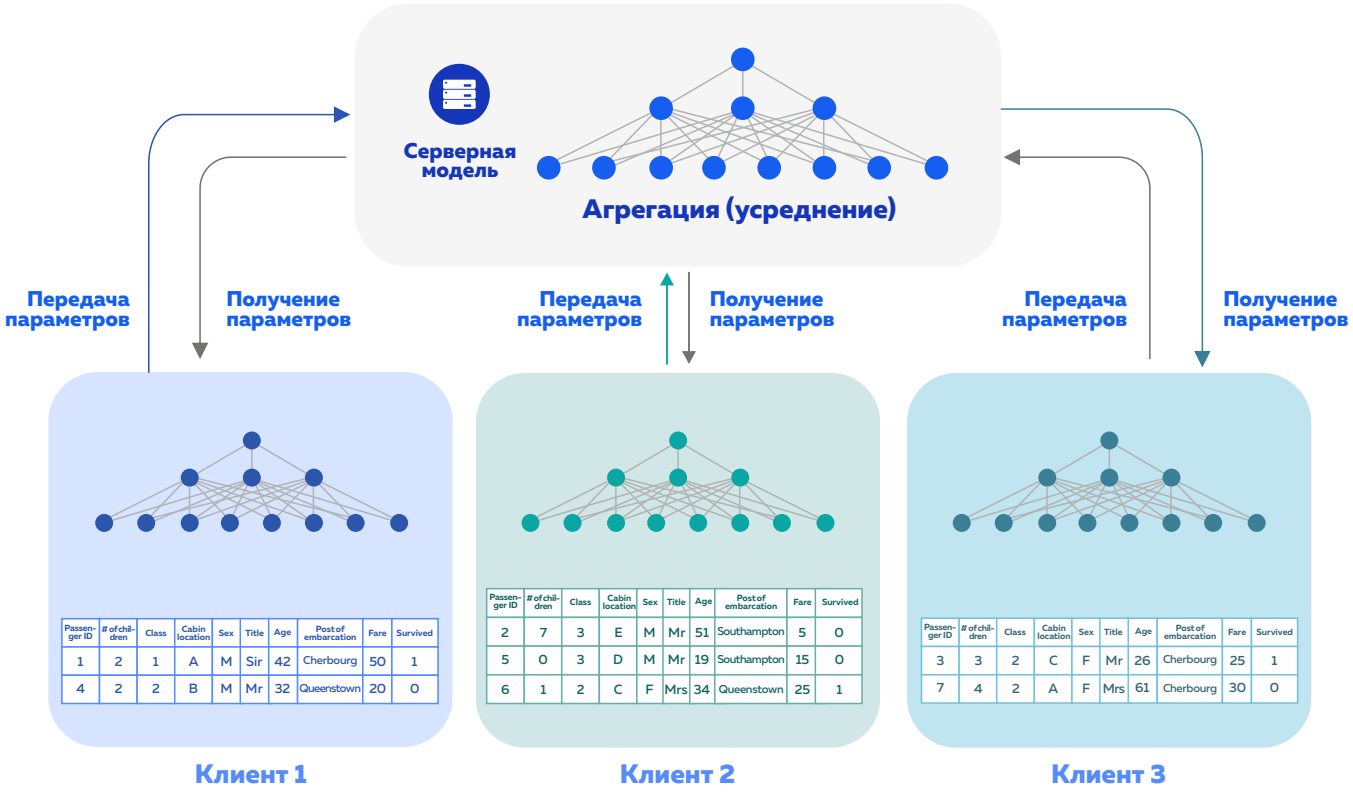


Рис. 2. Схема функционирования HFL

Важным этапом организации HFL является согласование архитектуры глобальной модели, процедуры предварительной обработки данных, соотношения локальных и глобальных раундов обучения и других гиперпараметров. На рис. 2 проиллюстрированы следующие принципы HFL:

- каждый клиент начинает обучение, используя одни и те же параметры модели;
- процедура локального обучения может варьироваться от набора нескольких батчей (пакетов данных) до нескольких эпох (полных проходов по всему набору обучающих данных), что, как правило, связано с объемами передаваемых параметров и пропускной способностью каналов связи;
- обновления моделей могут быть представлены либо непосредственно весовыми коэффициентами, либо градиентами (обеспечение защиты их значений будет рассмотрено в главе «Модель рисков»);
- роль управляющего сервера, осуществляющего оркестровку обучения, может быть возложена на одного из клиентов.

Единственным принципиально важным различием реализаций HFL, помимо архитектуры модели, может быть только способ агрегирования локальных обновлений. В настоящее время предложено значительное количество таких способов [1, 2, 3, 4, 5, 6], каждый из которых имеет свои особенности, достоинства и ограничения. Они по-разному сказываются на скорости обучения модели и метриках качества.

Стоит отметить наблюдаемый сегодня широкий интерес к разработке библиотечных реализаций HFL. Наиболее распространенные и функциональные среди них — Flower, PySyft и TFF.

Вертикальное федеративное обучение

Вертикальное федеративное обучение применимо в случае, когда клиенты располагают разными признаками совпадающих сущностей, то есть когда между данными клиентов можно провести воображаемую вертикальную границу (см. рис. 3). Для обучения модели в этом случае необходимым и достаточным условием является наличие целевого признака (разметки) только у одного из участников обучения. Например, банк и интернет-магазин могут иметь информацию о тех же пользователях, но один хранит данные о финансовых транзакциях, а другой — о покупательских предпочтениях. При реализации VFL-подхода они могут обучить модель для оценки кредитного риска клиентов банка (для этого используется целевой признак банка) или модель для более точных рекомендаций товара клиентам интернет-магазина (для этого берется его целевой признак). Кроме того, они могут обучить обе модели в интересах каждого.

Рис. 3. Схема разделения данных при VFL

VFL, являющееся развитием идеи распределенного обучения, значительно отличается от HFL как в части своей организации, так и в плане последующей эксплуатации обученных моделей. В научных публикациях сейчас наблюдается активный интерес именно к этому направлению FL. Причина кроется, вероятно, в том, что VFL предоставляет компаниям и организациям способы взаимодействия, обеспечивающие использование локальных данных для совместного машинного обучения, тогда как без применения этой технологии в некоторых случаях (например, при отсутствии разметки) извлечение пользы из накопленных данных не представляется возможным.

Например, китайский Webank использует[2] VFL для сотрудничества с агентством по взысканию долгов. На основе данных о транзакциях и банковских операциях, предоставляемых Webank, и информации о выставленных счетах и имеющихся задолженностях от агентства удалось получить более точную модель оценки кредитного риска без передачи друг другу исходных данных и с сохранением конфиденциальных данных клиентов.

В ходе обучения каждый из FL-клиентов использует принадлежащий ему датасет в качестве входных данных локальной части модели, результаты промежуточной обработки которой передаются на центральный сервер. Здесь поступающие от FL-клиентов данные некоторым образом объединяются, после чего поступают на вход финального отрезка модели, который может быть как обучаемым преобразованием, так и фиксированной функцией, например сигмоидой. Сервер, располагая таргетными метками и выходами модели, вычисляет значение функции потерь и градиенты. Алгоритм обратного распространения ошибки обеспечивает передачу градиентов клиентам, что в совокупности позволяет оптимизировать все части модели.

[2] WeBank. Utilization of FATE in risk management of credit in small and micro enterprises. – 2019.

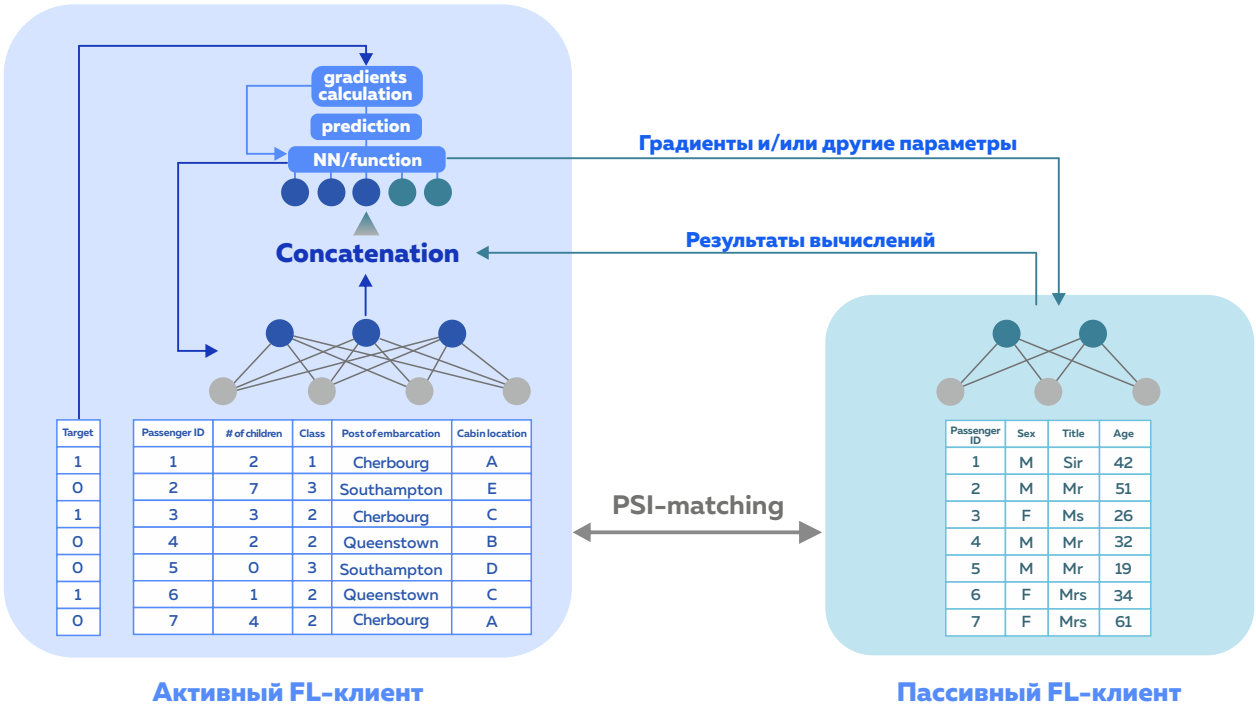


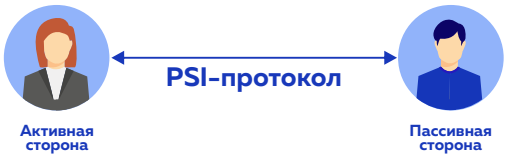
Рис. 4. Схема функционирования VFL

На практике реализовать VFL с доверенным сервером довольно сложно из-за того, что, во-первых, в этом случае на сервер необходимо в открытом виде передать значения целевого признака, что не соответствует концепции FL, и, во-вторых, нужно убедить участников обучения в том, что некоторой третьей стороне можно доверять, а это проблематично. Поэтому прикладное VFL, как правило, реализуется без участия выделенного доверенного сервера — его функции на себя берет один из FL-клиентов, обладающий целевым признаком. Такого клиента принято называть активным, а остальных — пассивными.

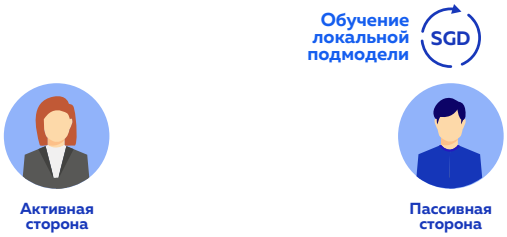
Порядок взаимодействия двух FL-клиентов в процессе обучения, схематично изображенного на рис. 4, можно представить в виде следующей последовательности шагов:

КОНФИДЕНЦИАЛЬНЫЕ ВЫЧИСЛЕНИЯ И ДОВЕРЕННЫЕ СРЕДЫ ИСПОЛНЕНИЯ
Федеративное обучение (Federated learning)

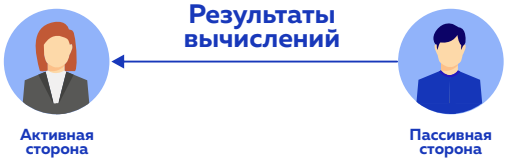
Шаг 1. Совместное обучение начинается с процедуры согласования данных: необходимо осуществить пересечение данных клиентов, чтобы выявить сущности с полными наборами признаков, например, с использованием PSI-методов, рассмотренных в предыдущей статье.



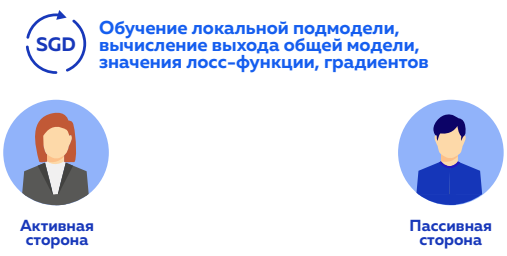
Шаг 2. Для каждого поданного на вход пассивными сторонами обучающего примера вычисляются промежуточные выходы локальных частей модели.



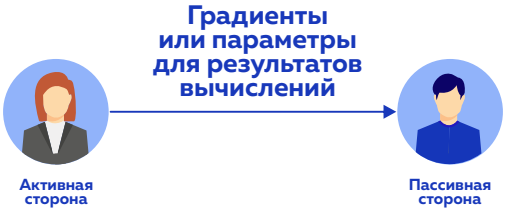
Шаг 3. Промежуточные выходы передаются активной стороне.



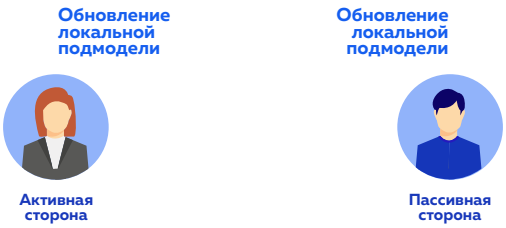
Шаг 4. Активная сторона вычисляет общий выход модели, производит расчет функции потерь и градиентов, после чего обновляет параметры модели.



Шаг 5. Основываясь на промежуточных выходах, активная сторона проводит вычисление градиентов, которые направляются соответствующим пассивным сторонам.



Шаг 6. Каждая сторона в соответствии с вычисляемыми ею градиентами обновляет параметры своей локальной модели.



Шаг 7. Шаги 2–6 повторяются до выполнения критерия остановки процедуры обучения.

Стоит подчеркнуть, что в случае использования VFL клиенты становятся взаимозависимыми, то есть при инференсе, как и при обучении, требуется кооперация всех сторон. Кроме того, применение VFL требует больших трудозатрат в ходе адаптации алгоритма обучения для каждого конкретного случая распределения признаков между сторонами и используемых архитектур, поэтому общедоступных фреймворков и библиотек с реализацией VFL крайне мало. На текущий момент наиболее известны FATE и NVFlare — фреймворки федеративного обучения, содержащие ограниченный перечень архитектур моделей, поддерживающих VFL.

Федеративное трансферное обучение

В 2009 году была предложена концепция переноса знаний — трансферного обучения[3] — между различными доменами данных. В рамках этой концепции исследовались подходы и методы, позволяющие адаптировать модели, обученные для решения задач, например, в банковском секторе в определенном регионе, для решения задач в других сегментах бизнеса — отраслевых или региональных. Логичным развитием концепции стало ее объединение с технологией федеративного обучения, давшее начало новому типу FL — федеративному трансферному обучению[4] (FTL). В отличие от HFL, где все участники имеют одинаковые признаки, и от VFL, где признаки различаются, но совпадают сущности, FTL позволяет обмениваться знаниями между участниками обучения даже при отсутствии общего глобального пространства признаков или сущностей.

Применимость такого подхода можно рассмотреть на примере наборов данных, собираемых предприятиями, имеющими близкий, но все же неодинаковый профиль деятельности. Из-за различий в характере бизнеса такие предприятия имеют лишь небольшое пересечение в пространстве признаков и, возможно, сущностей. Например, используя FTL, могут сотрудничать страховая компания и автосервис: у страховой компании имеется история страховых случаев, а у автосервиса — данные о ремонтах автомобилей. Несмотря на различие большинства признаков и сущностей, их данные можно объединить и построить более точную модель прогнозирования аварий. Аналогичным образом, используя FTL, можно объединять данные предприятий, расположенных в разных регионах мира.

Чтобы адаптировать модели, обученные на одном наборе данных, для применения на данных из другого отраслевого или регионального домена, в FTL используются методы адаптации доменов (domain adaptation), переноса параметров (parameter transfer) и пр. Схематично принцип разделения данных между участниками FTL представлен на рис. 5.

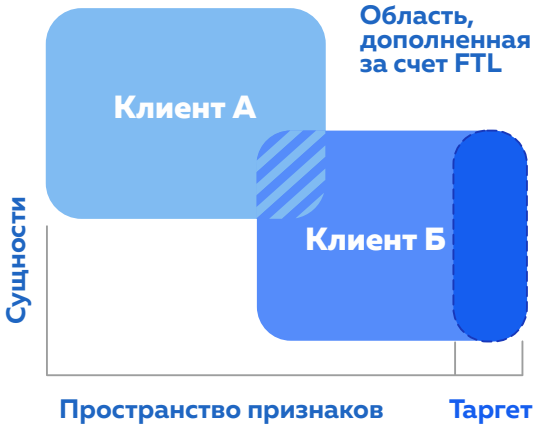


Рис. 5. Схема разделения данных при FTL

Пусть имеются два участника совместного обучения — А и Б, данные которых имеют лишь небольшое число пересекающихся признаков и сущностей (выделены пунктирной рамкой на рис. 1). Цель FTL — осуществить перенос знаний, накопленных в ходе обучения модели на данных участника А, с помощью информации, извлекаемой из пересекающихся данных. Другими словами, цель — адаптировать модель, обученную на образцах и признаках одного домена, для корректной работы с данными другого (целевого) домена. Таким образом, FTL позволяет клиенту Б обогащать свою модель знаниями, полученными клиентом А.

Технология FTL обладает большим потенциалом, расширяя возможности HFL и дополняя VFL. Уже сейчас на ее основе созданы алгоритмы обучения на медицинских данных моделей для классификации электроэнцефалограмм мозга [7] и обучения моделей автономного управления автомобилями [8]. В работе [9] предлагается фреймворк FedSteg, предназначенный для обучения с применением FTL модели, детектирующей наличие сокрытой стеганографическими методами информации в изображениях. Однако, учитывая очень высокую сложность технологии, необходимо тщательно настраивать и анализировать данные, передаваемые участниками друг другу, чтобы предотвращать утечки конфиденциальной информации.

[3] S. J. Panand, Q. Yang, A survey on transfer learning, IEEE Transactions on knowledge and data engineering, vol. 22, no. 10, pp. 1345–1359, 2009.

[4] Y. Liu, Y. Kang, C. Xing, T. Chen and Q. Yang, A secure federated transfer learning framework, IEEE Intelligent Systems, vol. 35, no. 4, pp. 70–82, 2020. https://arxiv.org/abs/1812.03337.

Основные отличия HFL, VFL и FTL

	HFL	VFL	FTL
Разделение данных	Пространство сущностей	Пространство признаков	Пространства сущностей и признаков
Сценарий обучения	Cross-device, cross-silo	Cross-silo	Более распространен cross-silo
Циркулирующие данные	Параметры модели	Промежуточные результаты вычислений	Промежуточные результаты вычислений
Тип локальной информации	Данные	Модели клиентов, данные	Модели клиентов, данные
Сложность адаптации алгоритма обучения	Средняя	Высокая	Очень высокая
Участник получает в итоге	Глобальную модель	Локальную часть модели	Локальную модель
Самостоятельное использование модели	Возможно	Невозможно	Возможно

Основное отличие VFL от HFL и FTL состоит в том, что инференс VFL-обученной модели также необходимо выполнять коллективно, что влечет за собой как дальнейшую взаимозависимость сторон, так и необходимость обеспечения конфиденциальности поступающей на инференс информации.

ПРЕИМУЩЕСТВА И НЕДОСТАТКИ ТЕХНОЛОГИИ

FL помогает сформировать среду для распределенного машинного обучения, обеспечивая следующие преимущества:

- **Конфиденциальность.** Благодаря тому, что отсутствует передача исходных данных, FL минимизирует риск утечек или несанкционированного доступа напрямую к конфиденциальной информации.
- **Масштабируемость.** Благодаря децентрализации вычислений FL позволяет эффективно обрабатывать большие объемы данных и расширять обучение на большое число клиентов или устройств без значительного увеличения нагрузки на сети или вычислительные ресурсы.
- **Эффективное использование ресурсов.** FL распараллеливает обучение глобальной модели без необходимости оснащения агрегирующей стороны значительными вычислительными мощностями.
- **Снижение эффекта дрейфа данных.** Путь данных от источника до модели становится короче — это уменьшает вероятность их устаревания или искажения.

Однако, как и любая другая технология, FL не лишено недостатков. Перечислим наиболее существенные из них:

- **Сложность координации.** Для управления процессом обучения с участием многих клиентов требуется сложная система координации и согласования — это может затруднить развертывание и поддержку системы.
- **Вычислительные ограничения.** Вычислительные ресурсы клиентов или пользовательских устройств могут быть ограничены — это затрудняет обучение сложных моделей и вынуждает проводить дополнительную оптимизацию алгоритмов.
- **Необходимость кооперации.** Не так просто бывает найти владельцев данных, желающих и готовых совместно решать практическую задачу.
- **Необходимость согласования данных.** Различия в методах построения данных и способах формирования идентификаторов сущностей могут привести к рассогласованности локальных моделей или потере глобальной моделью своей общности.
- **Угрозы безопасности.** FL требует повышенного внимания к защите от киберугроз и мошенничества из-за риска атак на отдельные устройства или серверы, хранящие данные либо обновления моделей, в результате чего злоумышленники, используя градиентную информацию, могут для некоторых архитектур моделей получать сведения об исходных данных.

СЦЕНАРИИ ПРИМЕНЕНИЯ FL

Благодаря тому, что FL позволяет использовать конфиденциальные и распределенные данные для построения ML-моделей, не передавая исходные данные в централизованные хранилища, появляется возможность задействовать в обучении неразмеченные и плохо структурированные данные. Это делает FL востребованным в отраслях, где важно соблюдать приватность и защиту информации. Приведем наиболее перспективные области применения FL.

Финансовый сектор

FL имеет значительный потенциал в финансовом секторе, особенно в условиях ужесточения регулирования и роста требований к защите данных. Финансовые организации, объединяясь друг с другом или с телеком-операторами, страховыми и ретейловыми компаниями, могут использовать федеративное обучение для создания моделей, решающих весьма актуальные задачи:

- **Выявление мошенничества и фрод-активности** FL позволяет банкам объединять информацию о подозрительных операциях, не раскрывая клиентские данные. Анализ транзакционных шаблонов по данным нескольких организаций позволяет выявлять мошеннические операции существенно эффективнее, чем при использовании данных только одного банка.

- **Улучшение кредитного скоринга.** Используя данные из разных регионов, банки могут совместно разрабатывать модели оценки кредитоспособности, не раскрывая информацию о конкретных клиентах. И если традиционные системы кредитного скоринга анализируют историю платежей клиентов в какой-то одной организации, то FL позволяет учитывать информацию из нескольких организаций (банков, страховых компаний, розничных сетей, интернет-магазинов), не нарушая конфиденциальности клиентов. Кроме того, FL может помочь страховым компаниям и банкам совместно анализировать риски заемщиков, давая возможность предложить надежным клиентам более привлекательные процентные ставки.

- **Повышение качества управления рисками.** FL может использоваться для анализа рисков на основе данных из разных филиалов или организаций-партнеров. В частности, банки и хедж-фонды могут применять FL для совместного обучения моделей прогнозирования рыночных рисков, опираясь на данные из разных источников.

- **Персонализация финансовых услуг.** FL позволяет анализировать поведение клиентов, обеспечивая обучение моделей непосредственно на их пользовательских устройствах. Это дает возможность предлагать персонализированные продукты без ущерба для конфиденциальности чувствительных данных.

Медицина

В области медицины конфиденциальность данных, их распределенность и необходимость их совместного использования являются ключевыми факторами, влияющими на внедрение и эксплуатацию информационных систем. Применение FL помогает воплотить в жизнь принципиально новые подходы к анализу данных, разработке моделей и принятию решений в здравоохранении.

- **Улучшение диагностики и прогнозирования заболеваний.** FL позволяет объединять данные из разных клиник, больниц и исследовательских центров по всему миру, что обеспечивает возможность строить модели, помогающие более точно выявлять опухоли при анализе медицинских изображений (рентгеновских снимков, МРТ, КТ), прогнозировать сердечно-сосудистые заболевания, диагностировать редкие заболевания и патологии, предсказывать осложнения или результаты лечения с учетом региональных, генетических, экологических и других факторов.

- **Мониторинг пациентов.** FL может использоваться для анализа состояния здоровья в режиме реального времени на основе данных с носимых устройств (фитнес-браслеты, умные часы, медицинские сенсоры). Непрерывный мониторинг пациентов помогает выявлять признаки развития инсультов, инфарктов и диабетических кризов, обнаруживать аномалии в показателях здоровья (пульс, давление, уровень сахара в крови), формировать персонализированные рекомендации по физической активности и питанию и пр.

Мобильные устройства и персонализированные модели

FL позволяет обучать модели на данных, которые остаются на устройствах пользователей, что открывает новые возможности для создания безопасных, энергоэффективных, адаптивных систем ИИ. Мы видим следующие основные направления применения FL:

- **Совершенствование клавиатурных приложений и голосовых помощников.** FL может использоваться для улучшения предсказания текста и распознавания речи с учетом индивидуальных особенностей пользователей (акцент, стиль и темп речи). Элементы FL уже используют такие мобильные приложения для управления клавиатурами, как Gboard и SwiftKey, создающие с помощью FL персонализированные рекомендации без нарушения конфиденциальности пользователей.

- **Компьютерное зрение на устройствах.** FL позволит мобильным камерам лучше адаптироваться к условиям съемки, точнее распознавать лица, жесты и объекты без передачи изображений на сервер.

- **Оптимизация энергопотребления и работы батареи.** FL поможет обучить модели, работающие на устройстве и анализирующие поведение пользователя, динамически управляя расходом заряда аккумулятора и адаптируя алгоритмы энергосбережения под особенности работы конкретного устройства.

- **Рекомендательные системы.** Использовать FL для персонализации рекомендаций могут соцсети, новостные агрегаторы, адаптивные рекламные алгоритмы, умные музыкальные плейлисты и приложения для потокового вещания.

- **Адаптивные обучающие системы.** Приложения для изучения языков и образовательные платформы могут адаптировать свои курсы с учетом показателей индивидуального прогресса и стиля обучения конкретного пользователя.

- **Здоровье и фитнес.** Приложения для мониторинга здоровья могут обучать модели на основе данных с датчиков (например, пульс, шаги), чтобы затем предоставлять владельцам устройств персонализированные рекомендации.

- **Навигация и карты.** FL может использоваться для оптимизации маршрутов и прогнозов трафика на основе данных с устройств множества пользователей.

Интернет вещей

Нельзя не отметить широкие перспективы FL в области интернета вещей (IoT), где устройства генерируют огромные объемы данных, но их передача на централизованные серверы может быть затруднена из-за недостаточной пропускной способности, задержек, а также требований к конфиденциальности. Основными направлениями применения FL в IoT могут стать следующие:

- **Обеспечение конфиденциальности данных.** Устройства IoT зачастую собирают чувствительные данные и персональную информацию — видео лиц, попадающих в умные камеры, запись голосов в умных колонках, данные о здоровье и поведении пользователей, показатели промышленных процессов и пр. Федеративное обучение позволяет обучать модели, не передавая сырые данные на центральные серверы — это снижает риски утечек и нарушений конфиденциальности, что особенно важно в свете ужесточения регуляторных требований.

- **Эффективное использование ресурсов и оптимизация работы умных устройств.** FL помогает более эффективно распределить вычислительную нагрузку между компонентами IoT, зачастую имеющими весьма ограниченные вычислительные ресурсы и энергопотребление, и мощными серверами, что снижает планку требований к централизованным вычислительным ресурсам. В свою очередь, современные методы оптимизации (сжатие моделей, квантование и пр.) обеспечивают возможность адаптации FL, что позволяет включать в контур обучения даже маломощные устройства.

- **Снижение задержек и нагрузки на сеть.** Передача больших объемов данных с устройств IoT в облако может вызывать задержки и перегружать сеть. FL минимизирует передачу данных, так как на сервер отправляются только обновления модели, а не сырые данные. Это особенно важно для приложений, работающих в реальном времени, таких как автономные транспортные системы или комплексы промышленной автоматизации.

- **Масштабируемость.** FL позволяет эффективно масштабировать обучение моделей, охватывая IoT-сети с миллионами устройств, поскольку каждое из них может участвовать в процессе обучения независимо от других. Кроме того, алгоритмы FL могут адаптироваться к динамически меняющейся сети устройств и создавать самообучающиеся IoT-системы, где устройства делятся друг с другом новыми знаниями.

- **Кибербезопасность.** IoT-системы подвержены кибератакам, особенно это касается централизованных решений. FL позволит выявлять аномалии в сетевом трафике, распределяя анализ угроз между устройствами. Кроме того, FL поможет повысить киберустойчивость системы (так как взлом одного узла не приведет к компрометации всей модели)

и защитить устройства от вредоносного ПО, используя распознавание подозрительного поведения без передачи конфиденциальных данных вовне. Например, умные маршрутизаторы смогут коллективно обучаться на сигнатурах кибератак, предотвращая угрозы практически в реальном времени.

Автономные автомобили

Автономные автомобили собирают большие объемы данных, включая информацию о местоположении, погоде, поведении водителей и пешеходов, дорожных условиях, телеметрию установленного на автомобиль оборудования и т. д. Эти данные могут носить в том числе и личный, конфиденциальный характер, поэтому их передача на централизованные серверы справедливо вызовет опасения у участников дорожного движения.

Федеративное обучение позволит реализовать принципиально новые возможности:

- обучать модели распознавания объектов (знаков, пешеходов, велосипедистов) непосредственно на устройствах (автомобилях), используя данные из разных регионов, оптимизировать принятие решений в реальном времени на основе локального и глобального опыта, адаптировать модели к локальным особенностям движения (например, к различиям в ПДД или манере вождения в разных странах);
- непрерывно обучаться на новых данных, собираемых в реальном времени, и параллельно отправлять и получать обновления для модели, что ускоряет процесс внедрения улучшений;
- распределять вычислительную нагрузку между устройствами, что по мере роста числа автомобилей сделает процесс обучения более масштабируемым и экономически эффективным.

Рекламные технологии и розничная торговля

Федеративное обучение открывает новые горизонты в области рекламных технологий (AdTech) и розничной торговли, что предоставит компаниям возможность анализировать пользовательское поведение, повышать эффективность маркетинговых кампаний и улучшать клиентский опыт, сохраняя конфиденциальность данных. Вот далеко не полный список задач, которые можно решать с помощью FL:

- **Персонализированная реклама.** FL позволяет обучать рекламные модели на основе данных с пользовательских устройств (смартфонов, ПК, смарт-ТВ, умных колонок) — это улучшает релевантность рекламы и повышает конверсию.
- **Адаптивность рекламы.** Системы машинного обучения смогут локально анализировать взаимодействие пользователей с рекламой и адаптировать ее в реальном времени, используя данные, полученные от множества рекламодателей и партнеров.
- **Анализ данных в реальном времени.** Розничные сети могут применять FL для анализа данных с кассовых терминалов, мобильных приложений и IoT-устройств (например, умных полок) без передачи данных в централизованные системы. В частности, модели могут использоваться для прогнозирования спроса на товары, оптимизации ассортимента и цепочек поставок, инвентаризации и управления запасами.
- **Борьба с мошенничеством и возвратами.** FL поможет выявлять аномальные транзакции и потенциальные случаи мошенничества и злоупотребления программами лояльности, предотвращать фродовые транзакции в онлайн-магазинах, анализировать схемы возвратов и подозрительных покупок без централизованного сбора клиентских данных.

Государственные учреждения и органы власти

Государственные учреждения и органы власти работают с огромными объемами чувствительных данных, включающими информацию о гражданах, экономике, безопасности и здравоохранении. Федеративное обучение позволит значительно повысить эффективность их работы, что обеспечит конфиденциальность данных и улучшит аналитические возможности по целому ряду направлений:

- **Национальная безопасность и деятельность правоохранительных органов.** FL может применяться для анализа угроз, предотвращения преступлений и координации силовых структур, проведения совместного анализа подозрительных финансовых операций налоговыми службами и спецслужбами, выявления преступных схем с использованием распределенных данных (например, с целью анализа кибератак на основе данных из разных стран).

- **Здравоохранение и эпидемиологический мониторинг.** FL позволит государственным органам обрабатывать медицинские данные из множества клиник и регионов с целью обучения моделей для раннего выявления эпидемий, а также ранней диагностики заболеваний за счет анализа большего объема медицинских данных, сохраняя конфиденциальность пациентов.

- **Финансовый сектор и борьба с мошенничеством.** Государственные структуры могут использовать FL для создания моделей контроля за налоговыми операциями с целью выявления налоговых махинаций путем анализа данных разных организаций, отслеживания банковских транзакций и борьбы с отмыванием денег, вскрывая коррупционные схемы и факты незаконного финансирования.

МОДЕЛЬ РИСКОВ

Концепция федеративного обучения позволяет свести к минимуму риск непосредственной утечки данных, однако при использовании FL возникают новые угрозы, анализ которых требует особого внимания при построении модели рисков.

Риск компрометации данных

Несмотря на то, что FL предлагает среду для распределенного обучения с минимизацией передачи непосредственно конфиденциальных данных, существует риск утечки информации через промежуточные результаты обучения в ходе атак, которые могут осуществляться как внешним злоумышленником, так и одним из участников. Например, в работах [10, 11, 12] демонстрируются уязвимости FL-схем, позволяющие получать информацию об исходных данных на основе анализа градиентов и параметров модели, передаваемых участниками друг другу, — утечки здесь возможны в случае компрометации каналов связи и центрального сервера.

Кроме того, в отличие от HFL, вертикальное федеративное обучение предполагает взаимодействие не доверяющих друг другу сторон не только в процессе обучения модели, но и при проведении инференса. Данный факт также необходимо учитывать при построении модели рисков, поскольку недобросовестный участник может попытаться получить несанкционированный доступ к чувствительной информации других участников. В частности, в работах [13, 14] рассматриваются возможности восстановления значений целевого признака в процессе совместного инференса.

Противодействие:

- защита данных модели, передаваемых друг другу сторонами обучения, путем использования полного (Fully Homomorphic Encryption, FHE) или частичного (Partially Homomorphic Encryption, PHE) гомоморфного шифрования, дифференциальной приватности (Differential Privacy, DP) или иных методов защиты;

- защита каналов связи между участниками с применением криптографических методов.

Такой комбинированный подход позволяет реализовать преимущества распределенного обучения моделей FL и обеспечить его доказуемую стойкость к значительному числу возможных атак рассматриваемого вида.

Риск компрометации модели (отравление данных)

Внешний злоумышленник или недобросовестный участник обучения может целенаправленно обучать свою локальную модель на некорректных данных и отправлять искаженные параметры на агрегацию, чтобы ухудшить качество или изменить поведение глобальной модели. Например, атакующий может добавлять скрытые триггеры в данные так, чтобы в определенных случаях модель выдавала неверные прогнозы. Подобные атаки подробно рассмотрены в [15, 16].

Противодействие:

- включение механизмов аутентификации;
- использование методов обнаружения аномалий (Byzantine-Resilient Aggregation, Krum, Median), помогающих отфильтровывать подозрительные обновления моделей;
- внедрение механизмов регуляризации параметров и ограничения вклада каждого FL-клиента (Norm Clipping, Adaptive Federated Learning), снижающих влияние отдельных участников на глобальную модель.

Из-за риска компрометации модели технологию FL на практике применяют только в комбинации с другими техниками и методами защиты передаваемых между сторонами обучения параметров модели — FHE, PHE, DP и др. Такой подход позволяет реализовать преимущества распределенного обучения моделей FL и обеспечить его доказуемую стойкость к значительному числу возможных атак.

Стоит отметить, что применение таких частично гомоморфных криптоалгоритмов, как RSA, который гомоморфен относительно операции умножения, или схемы Paillier, гомоморфной относительно операции сложения, хотя и приводит к росту вычислительной и емкостной нагрузки процедуры обучения, не ускоряет его слишком сильно — с кратностью в три порядка и более, как в FHE-схемах.

Численная оценка рисков

Наиболее полные модели оценки рисков [17, 18] базируются на прогнозировании успешности атак на FL-модели. Хотя архитектура FL в основном гарантирует отсутствие прямого обмена данными, косвенные риски их восстановления через дополнительные источники данных все еще имеются.

Успешность атаки на данные и модель FL может быть оценена через вероятность по формуле Байеса:

$$P_{attack} = P(\text{успех}|\text{имеющиеся ресурсы, данные}) = \frac{P(\text{ресурсы}|\text{успех}) \times P(\text{успех}) \times Ф}{P(\text{ресурсы})}.$$

В этой формуле:

***P* (успех)** — базовая априорная вероятность успеха атаки на незащищенную систему FL, в большинстве случаев она может быть оценена как $P(\text{успех}) = \frac{\text{количество успешных атак}}{\text{общее количество попыток}}$;

***P* (ресурсы|успех)** — вероятность того, что злоумышленник обладает необходимыми ресурсами для достижения успеха за ограниченное время;

***P* (ресурсы)** — вероятность наличия ресурсов в целом;

Ф - фактор защиты, связанный с применением схемы гомоморфного шифрования или дифференциальной приватности.

Базовая априорная вероятность незащищенных систем FL оценивается на основе их симуляции или экспертным методом. Истинное значение P(успех) определить довольно сложно, так как оно зависит от множества факторов, среди которых одним из наиболее значимых является архитектура обучаемой модели. Диапазон допустимых значений оценки априорной вероятности довольно широк, но в большинстве случаев для практического использования в качестве оценки того, что будет в худшем случае, может использоваться значение 0,5, что подтверждается рядом исследований, например [18–21].

Расчет вероятности получения необходимых для успешной атаки ресурсов P(ресурсы|успех) можно оценить одним из двух способов. Первый — на основе сложности модели (здесь λ — константа масштабирования в диапазоне от 0,1 до 0,5):

$$P(\text{ресурсы}|\text{успех}) = e^{-\lambda \frac{\text{сложность модели}}{\text{доступные ресурсы}}}.$$

Второй способ — на основе оценок требований к ресурсам:

$$P(\text{ресурсы}|\text{успех}) = min\ (1, \frac{\text{доступные ресурсы}}{\text{требуемые ресурсы}}).$$

При отсутствии должных симуляций или надежных данных о требованиях к ресурсам следует принимать значение *P (ресурсы|успех) ≈ 1.*

Оценка наличия ресурсов в целом (в индустрии, по рынку) — *P* (ресурсы) — определяется с учетом доступа различных участников информационного обмена и потенциальных злоумышленников к информации для обмена:

$$P(\text{ресурсы}) = \frac{\text{количество организаций с ресурсами}}{\text{общее количество потенциальных атакующих}} = e^{-\frac{\text{стоимость ресурсов}}{\text{средний бюджет атакующего}}}$$

При отсутствии надежных данных можно руководствоваться оценками для различных типов атакующих (их моделирование проведено Ассоциацией больших данных), приведенными в таблице ниже.

Тип атакующего	Доступные ресурсы	*P* (ресурсы)
Любитель	Низкие	0,01–0,05
Хакерская группа	Средние	0,1–0,3
Корпоративный конкурент	Высокие	0,3–0,6
Государственный институт	Очень высокие	0,7–0,95

Например, для атаки, нацеленной на восстановление данных в HFL-системе с 10 участниками и моделью средней сложности, значения вероятностей будут такими:

P (ресурсы|успех) ≈ 0,4 (нужны значительные, но не экстремальные ресурсы);

P (ресурсы) ≈ 0,2 (такие ресурсы есть у примерно 20% потенциальных атакующих);

P (успех) ≈ 0,25 (базовая вероятность для этого типа атаки).

Для незащищенной схемы *(Ф=1)* получаем довольно высокую оценку:

$$P_{attack} \approx \frac{0{,}4 \times 0{,}25}{0{,}2} \approx 0{,}5$$

Однако при использовании защиты на основе гомоморфного шифрования или дифференциальной приватности получаем следующие величины:

- в случае с гомоморфным шифрованием

$$\Phi = P(kbreak) = C \times T \times 2^{-L}$$

— отражает вероятность взлома криптографического ключа, где C — вычислительные ресурсы атакующего, T — время, доступное для атаки, L — уровень криптостойкости;

- при использовании дифференциальной приватности $\Phi = e^{-\epsilon}$, где ϵ — параметр дифференциальной приватности (бюджет приватности).

Оценка применения схемы с гомоморфным шифрованием, например, для схемы Пайе с уровнем криптостойкости $L = 112$ (длина модуля 2048 битов), стандартным временным ограничением $T = 1$ год $\approx 3{,}15 * 10^7$ сек и доступными ресурсами $C \approx 10^{12}$ операций в секунду составляет:

$$\Phi = C \times T \times 2^{-L} \leq 3{,}15 \times 10^{-14} \to 0.$$

Для защиты методом дифференциальной приватности оценка скромнее: при $\epsilon = 1$ она составит

$$\Phi = e^{-1} \leq 0{,}37$$

Примечание: применение гомоморфного шифрования значительно усложняет общую схему FL, а дифференциальной приватности — ухудшает качественные характеристики модели в случае, если подбор значения бюджета приватности окажется несбалансированным.

Дополнительные риски

Также в модель рисков для систем федеративного обучения иногда включают следующие риски:

- **Риск асимметрии данных.** Данные участников FL могут сильно различаться по качеству и количеству, что чревато проблемами с генерализацией и обучением модели. Разные устройства или организации могут иметь разные данные — это может нарушить синхронность и эффективность обучения.

- **Риск рассинхронизации.** В ходе обмена параметрами модели FL важно обеспечить соблюдение синхронности и периодичности. Проблемы с коммуникацией участников могут замедлить обучение или даже полностью его остановить.

- **Риск ошибок и сбоев в системе.** Неисправности в устройствах или сбои в коммуникационных каналах могут привести к потере данных или некорректным обновлениям, что отразится на итоговой модели. Особенно это важно при работе с устройствами в реальном времени, когда сбои могут быть частыми.

- **Риски юридического характера.** Если участники FL имеют разную государственную принадлежность, различия в законодательных и этических нормах, могут возникнуть правовые препятствия для использования FL.

Поскольку все перечисленные риски вытекают из независимых событий, используется мультипликативный подход и формула отказа от риска:

$$P_{total} = 1 - \prod_{i=1}^{n} (1 - P_i),$$

где P_i - отдельные вероятности, такие как P_{attack}, $P_{asymmetry}$ и т. п.

ЮРИДИЧЕСКАЯ ИНТЕРПРЕТАЦИЯ ТЕХНОЛОГИИ

С юридической точки зрения технология федеративного обучения позволяет исключить передачу участниками друг другу исходных данных, включая персональные данные (ПДн). Участникам обучения нет необходимости получать согласие на обработку ПДн или обеспечивать иное правовое основание для обработки ПДн, операторами которых выступают другие участники. По этой причине технология феде-

ративного обучения относится к технологиям защищенной обработки данных (Privacy-Enhancing Technologies).

Важно отметить, что если речь идет об обучении моделей искусственного интеллекта на основе персональных данных, то важно обеспечить правомерность использования таких данных компаниями-участниками, включая наличие необходимого правового основания, и защитить их от рисков возможных утечек. В отношении основания обработки ПДн в целях обучения моделей ИИ существуют разные точки зрения. На текущий момент считается, что по умолчанию в качестве такого основания должно выступать согласие субъекта ПДн.

Что касается утечки ПДн, то она может возникнуть вследствие изъятия исходных данных, на которых производилось обучение, из обученной модели ИИ или их реконструкции при атаках через доступ к промежуточным результатам обучения, о чем шла речь выше. Чтобы минимизировать эти риски, нужно, наряду с использованием технологии FL, применять и иные технологии защищенной обработки данных, исключающие или существенно снижающие эти риски.

ПЕРСПЕКТИВЫ ПРИМЕНИМОСТИ ТЕХНОЛОГИИ FL

Общие тезисы о применимости

Технология FL уже используется крупнейшими мировыми компаниями в таких отраслевых секторах, как финансы и телекоммуникации. Высок потенциал использования FL в здравоохранении, образовании, энергетике, рекламе и маркетинге, системах интернета вещей и пр.

На рынке представлены как фреймворки с открытым кодом (Flower, NVFlare, TFF, FATE, PySyft и др.), так и коммерческие продукты (Guardora, Scaleout). FL легко масштабируется на практически неограниченное число устройств.

FL относится к программным методам обеспечения конфиденциальности данных. При его использовании данные остаются на устройствах или в локальных системах. Каналы связи между участниками обучения и передаваемая по этим каналам информация, как правило, дополнительно защищаются с применением методов с доказуемой стойкостью к атакам внешних злоумышленников и злонамеренных участников. Благодаря этому FL может применяться в задачах с повышенными требованиями к безопасности данных (например, в государственном управлении).

Общий прогноз. В условиях растущих требований к конфиденциальности данных и их безопасности количество областей применения федеративного обучения, сценариев и кейсов с его использованием будет расти как в мире, так и в России. Каждый новый кейс повышает общий интерес и доверие к этой технологии.

Направления дальнейших исследований

Развитие алгоритмов: оптимизация моделей и методов обучения для работы на устройствах с ограниченными ресурсами и низким энергопотреблением, внедрение систем интеллектуальной синхронизации участников.

Улучшение технологий защиты: широкое внедрение алгоритмов для выявления злонамеренных действий участников обучения и механизмов DP и PHE/FHE для предотвращения утечки данных.

Стандартизация и регламентирование: разработка общих внутригосударственных и международных стандартов и протоколов FL, стандартов безопасности обученных моделей и их использования.

ВЫВОДЫ

Технология машинного федеративного обучения может найти широкое применение в решении множества отраслевых задач (см. п. 4). Более того, в некоторых случаях альтернативных FL способов решения либо вовсе не существует, либо их трудно реализовать (например, при обучении медицинских моделей организациями из разных государств).

Сторонам, участвующим в федеративном обучении, нельзя забывать о том, что отсутствие передачи исходных данных не является гарантией их конфиденциальности. Наличие обшир-

ного списка угроз требует комплексного подхода к обеспечению конфиденциальности данных во время их использования в обучении, включающего в себя использование криптографических методов, а также внедрение механизмов для выявления и блокировки подозрительных активностей других участников.

Для внедрения эффективных решений необходимы компромиссные сценарии, сохраняющие практическую ценность метода и безопасность данных. Анализ этих подходов позволяет выявить их потенциал для создания безопасных и масштабируемых решений, способных выполнять сложные задачи анализа данных, сохраняя высокие стандарты защиты конфиденциальности и безопасности.

АВТОРЫ ДОКЛАДА

АЛЕКСЕЙ НЕЙМАН
Исполнительный директор
Ассоциации больших данных,
руководитель FIT Academy of Russia,
Master of Data Science, CDMP, PMP

ОЛЕГ ФАТЮХИН
Технический
руководитель проекта,
Guardora

МИХАИЛ ФАТЮХИН
Ведущий разработчик-
исследователь / криптограф,
Guardora

КИРИЛЛ ГРОШЕНКОВ
Ведущий исследователь,
Guardora

ВАЛЕРИЙ ХВАТОВ
Специалист
по кибербезопасности
и распределенным вычислениям,
технический директор
DGT Network

АЛЕКСАНДР ПАРТИН
Адвокат и партнер
Privacy Advocates,
соучредитель «РППА.Офис»,
сопредседатель
Privacy & Legal Innovation
кластера РАЭК, CIPP/E, CIPM

АЛЕКСЕЙ МУНТЯН
Генеральный директор Privacy Advocates,
внешний менеджер по защите данных
нескольких транснациональных холдингов,
соучредитель Regional Privacy
Professionals Association (RPPA.pro),
сопредседатель Privacy & Legal Innovation
и кластера РАЭК

ИСТОЧНИКИ

1. H. B. McMahan, E. Moore, D. Ramage. Federated learning of deep networks using model averaging. arXiv preprint arXiv:1602.05629, 2016.

2. T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar and V. Smith. Federated Optimization in Heterogeneous Networks. Proceedings of Machine Learning and Systems, Vol. 2, 429-450, 2020.

3. K. Bonawitz, V. Ivanov, B. Kreuter et al. Practical Secure Aggregation for Privacy-Preserving Machine Learning. Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, Dallas, 30 October-3 November 2017, 1175-1191, 2017.

4. P. Blanchard, R. Guerraoui, J. Stainer et al. Machine learning with adversaries: Byzantine tolerant gradient descent. In Advances in Neural Information Processing Systems, 119–129, 2017.

5. D. Yin, Y. Chen, R. Kannan and P. Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In Proceedings of the International Conference on Machine Learning, pp. 5650–5659, 2018.

6. J. Wang, Q. Liu, H. Liang, G. Joshi, H. V. Poor. Tackling the Objective Inconsistency Problem in Heterogeneous Federated Optimization. Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, December 6-12, 2020.

7 .C. Ju, D. Gao, R. Mane, B. Tan, Y. Liu, and C. Guan, Federated transfer learning for eeg signal classification, arXiv preprint arXiv:2004.12321, 2020.

8. X. Liang, Y. Liu, T. Chen, M. Liu, and Q. Yang, Federated transfer reinforcement learning for autonomous driving, arXiv preprint arXiv:1910.06001, 2019.

9. H. Yang, H. He, W. Zhang, and X. Cao, Fedsteg: A federated transfer learning framework for secure image steganalysis, IEEE Transactions on Network Science and Engineering, 2020.

10. B. Zhao, K. R. Mopuri, H. Bilen. Improved Deep Leakage from Gradients. arXiv preprint arXiv: 1906.08935, 2019.

11 .Y. Song, Z. Wang and E. Zuazua. Approximate and Weighted Data Reconstruction Attack in Federated Learning. arXiv preprint arXiv:2308.06822, 2023.

12 .Z. Wang, Z. Chang, J. Hu, X. Pang, J. Du, Y. Chen, K. Ren. Breaking Secure Aggregation: Label Leakage from Aggregated Gradients in Federated Learning. IEEE INFOCOM 2024, Vancouver, BC, Canada, 20-23 May 2024, 2024.

13. C. Fu, X. Zhang, S. Ji, J. Chen, J. Wu, S. Guo, J. Zhou, A. X. Liu and T. Wang. Label inference attacks against vertical federated learning. USENIX Security 22, Boston, MA: USENIX Association, Aug. 2022, 2022.

14. O. Li, J. Sun, X. Yang, W. Gao, H. Zhang, J. Xie, V. Smith and C. Wang. Label leakage and protection in two-party split learning. arXiv preprint arXiv:2102.08504, 2021.

15. E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, V. Shmatikov. How To Backdoor Federated Learning. arXiv preprint arXiv:1807.00459, 2020.

ИСТОЧНИКИ

16. C. Fung, C.J.M. Yoon, I. Beschastnikh. Mitigating Sybils in Federated Learning Poisoning. arXiv preprint arXiv:1808.04866, 2018.

17. Gong, Haimei, Liangjun Jiang, Xiaoyang Liu, Yuanqi Wang, Omary Gastro, Lei Wang, Ke Zhang, and Zhen Guo. "Gradient Leakage Attacks in Federated Learning." Artificial Intelligence Review 56, no. 1 (October 1, 2023): 1337–74. https://doi.org/10.1007/s10462-023-10550-z

18. Shirvani, Ghazaleh, Saeid Ghasemshirazi, and Behzad Beigzadeh. Federated Learning: Attacks, Defenses, Opportunities, and Challenges, 2024. https://doi.org/10.48550/arXiv.2403.06067

19. Song, Liwei, Reza Shokri, and Prateek Mittal. "Privacy Risks of Securing Machine Learning Models against Adversarial Examples." In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, 241–57, 2019. https://doi.org/10.1145/3319535.3354211

20. Zhu, Ligeng, Zhijian Liu, and Song Han. "Deep Leakage from Gradients." arXiv, December 19, 2019. https://doi.org/10.48550/arXiv.1906.08935

21. Rigaki, Maria, and Sebastian Garcia. "A Survey of Privacy Attacks in Machine Learning." ACM Computing Surveys 56, no. 4 (April 30, 2024): 1–34. https://doi.org/10.1145/3624010

Редакторы доклада

ЖАННА ПОКРОВСКАЯ
Руководитель пресс-службы
Ассоциации больших данных

АЛИЯ ТРУНАЕВА
PR-менеджер
Ассоциации больших данных

АССОЦИАЦИЯ БОЛЬШИХ ДАННЫХ

www.rubda.ru

Адрес: Москва,
Пресненская набережная, 10с2

+7 (495) 252-72-60
info@rubda.ru