УДК 004.85
https://doi.org/10.31854/2307-1303-2025-13-2-52-68
EDN: YHQXCK

Федеративное обучение с сохранением конфиденциальности: баланс между точностью и защитой данных в распределенном машинном обучении

Аль-Свейти Малик А. М. ✉, **Ким З. В.**, **Маршев Д. В.**

Санкт-Петербургский государственный университет телекоммуникаций им. проф. М. А. Бонч-Бруевича, Санкт-Петербург, 193232, Российская Федерация

Постановка задачи. *В условиях растущего объема чувствительных данных и ужесточения требований к их защите традиционные централизованные методы машинного обучения становятся неприемлемыми из-за рисков утечек и нарушения конфиденциальности. Особенно остро эта проблема стоит в таких сферах, как здравоохранение и финансы, где передача персональных данных на центральный сервер недопустима. Одним из перспективных решений является федеративное обучение, позволяющее обучать глобальные модели без передачи исходных данных, однако сохранение баланса между точностью модели и уровнем приватности остается ключевым вызовом.* ***Методы:*** *для решения задачи предложен подход, сочетающий алгоритм агрегации FedAvg с механизмами дифференциальной приватности, включая обрезку градиентов и добавление гауссовского шума на стороне клиентов. Экспериментальная валидация проведена на наборе данных MNIST с использованием сверточной нейронной сети при различных параметрах дифференциальной приватности.* ***Результаты:*** *при оптимальных настройках ($\sigma = 0{,}5$; $\varepsilon \approx 3$) достигнута точность 97,80 %, что лишь на 1 % уступает централизованному обучению (98,79 %). Безопасная агрегация с 10 клиентами за пять раундов показала точность 93,21 %. Анализ выявил четкую зависимость точности от параметров приватности, что позволяет гибко настраивать систему под конкретные требования.* ***Практическая значимость:*** *предложенная методика обеспечивает прозрачную и воспроизводимую оценку компромисса «точность – приватность», что делает ее применимой для внедрения в реальные системы с чувствительными данными; результаты могут быть использованы в качестве базы для адаптации федеративного обучения в медицинских, финансовых и других критически важных приложениях, где конфиденциальность является приоритетом.*

Ключевые слова: *федеративное обучение, дифференциальная приватность, машинное обучение, защита данных, компромисс «точность – приватность», безопасная агрегация*

Введение

В эпоху стремительного развития информационных технологий и цифровизации растет значение машинного обучения (МО) как инструмента для анализа и интерпретации больших объемов данных. Применение МО находит отражение в широком спектре областей – от диагностики заболеваний до прогнозирования финансовых рисков и обеспечения безопасности в киберпространстве.

Библиографическая ссылка на статью:
Аль-Свейти М. А. М., Ким З. В., Маршев Д. В. Федеративное обучение с сохранением конфиденциальности: баланс между точностью и защитой данных в распределенном машинном обучении // Информационные технологии и телекоммуникации. 2025. Т. 13. № 2. С. 52–68. DOI: 10.31854/2307-1303-2025-13-2-52-68. EDN: YHQXCK

Reference for citation:
Al Sweity M. A. M., Kim Z., Marshev D. Privacy-Preserving Federated Learning: Balancing Accuracy and Data Protection in Distributed Machine Learning // Telecom IT. 2025. Vol. 13. Iss. 2. PP. 52–68 (in Russian). DOI: 10.31854/2307-1303-2025-13-2-52-68. EDN: YHQXCK

При этом растет объем персональных и чувствительных данных, что порождает новые вызовы в области конфиденциальности и безопасности.

Традиционные централизованные подходы к МО предполагают передачу и хранение исходных данных на центральном сервере, что неизбежно повышает риски их утечки, несанкционированного доступа или нарушения конфиденциальности [1]. В качестве альтернативы было предложено федеративное обучение (ФО) – парадигма, при которой обучение глобальной модели (нейронной сети) происходит распределенно: локальные копии модели обучаются непосредственно на устройствах пользователей (клиентов), а исходные данные никогда не покидают их пределов. Вместо данных клиенты передают на сервер только обновления параметров своих локальных моделей (градиенты или веса), которые затем агрегируются для уточнения глобальной модели. Такой подход значительно снижает риски компрометации чувствительной информации, однако полностью не устраняет угрозы, связанные с возможной реконструкцией данных по передаваемым параметрам или атаками на этапе агрегации.

Несмотря на отсутствие прямого доступа к исходным данным, ФО остается уязвимым для ряда атак, таких как MIA (*аббр. от англ.* Membership Inference Attack – атаки на членство), MInvA (*аббр. от англ.* Model Inversion Attack – инверсия модели) и другие методы анализа локальных обновлений [2]. Для повышения уровня защиты конфиденциальности в ФО внедряются методы дифференциальной приватности (ДП). Концепция ДП предоставляет строгие математические гарантии, ограничивающие вероятность того, что выход модели позволит сделать вывод о присутствии или отсутствии конкретной записи в обучающем наборе данных (Dataset) [3]. В сочетании с ФО эти методы позволяют создавать системы, способные обеспечить высокую степень защиты данных, сохраняя при этом конкурентоспособную точность модели.

Цель данного исследования – интеграция механизмов ДП в архитектуру ФО и систематический анализ компромисса между точностью модели и уровнем конфиденциальности данных. Работа посвящена поиску баланса между защитой персональной информации и сохранением высокой эффективности обучения в распределенной среде, что особенно актуально при работе с чувствительными данными.

Несмотря на то что базовые принципы ФО с ДП уже описаны в ряде работ, практическая реализация и настройка параметров для достижения баланса между приватностью и точностью остаются сложной задачей, особенно в условиях ограниченных вычислительных ресурсов и неоднородности данных.

В отличие от теоретических исследований, данная работа фокусируется на эмпирической оценке влияния ключевых гиперпараметров (множитель шума, порог клиппинга, количество локальных эпох) на производительность модели в реалистичной настройке с использованием стандартной архитектуры сверточной нейронной сети (CNN, *аббр. от англ*. Convolutional Neural Network) и набора данных MNIST (*аббр. от англ*. Modified National Institute of Standards and Technology database – Модифицированная база данных Национального института стандартов и технологий).

Значимость данного исследования заключается в систематическом анализе компромисса «точность — приватность» и определении оптимальных параметров для практического применения механизмов ДП в системах ФО. Полученные результаты позволяют настраивать уровень защищенности под конкретные требования к конфиденциальности без значительной потери в точности модели, что делает их применимыми в реальных сценариях. Данная методология может служить основой для разработки и внедрения защищенных систем МО в промышленных условиях, особенно в сферах с повышенными требованиями к приватности – здравоохранении, финансах и телекоммуникациях.

Обзор литературы по теме исследования

Эволюция ФО. Первоначальные работы в области ФО заложили основу для создания алгоритма FedAvg (*аббр. от англ.* Federated Averaging – федеративное усреднение), который позволяет эффективно объединять локальные обновления, взвешивая вклад каждого клиента пропорционально объему его локального набора данных. При усреднении глобальных параметров модели сервер учитывает, сколько данных использовал каждый клиент для локального обучения; клиенты с бóльшим объемом данных оказывают пропорционально бóльшее влияние на финальную модель. Этот подход продемонстрировал возможность масштабирования распределенного обучения, что особенно важно в условиях, когда данные рассредоточены по множеству устройств. Базовые принципы ФО были расширены в более поздних исследованиях, посвященных оптимизации коммуникационных затрат, улучшению устойчивости моделей и адаптации к разнородности данных [4, 5].

ФО впервые было предложено как подход к обучению моделей на распределенных данных без их централизованного сбора, при этом основным механизмом выступало усреднение локальных обновлений, полученных с помощью SGD (*аббр. от англ.* Stochastic Gradient Descent – стохастический градиентный спуск) на клиентских устройствах: в [6] был разработан алгоритм FedAvg, позволивший существенно снизить объем передаваемых данных и сохранить приватность пользователей. Однако в реальных сценариях статистическая гетерогенность клиентских данных приводит к тому, что локальные модели «уходят» в разные стороны (явление, известное как дрейф локальных моделей), что, в свою очередь, замедляет или даже препятствует сходимости глобальной модели; в ответ на это были предложены методы контроля дисперсии градиентов, например, алгоритм SCAFFOLD, демонстрирующий устойчивость к неоднородности данных и улучшающий скорость сходимости в неблагоприятных условиях [7].

ДП в МО была впервые формализована в [3] как способ обеспечения математически обоснованных гарантий конфиденциальности. С появлением необходимости защищать данные при обучении глубоких нейронных сетей был предложен практический метод интеграции ДП в алгоритмы обучения, что позволило контролировать уровень риска утечки информации за счет добавления специально откалиброванного шума [8]. Эти методы нашли широкое применение

в коммерческих продуктах, таких как системы анализа данных крупных интернет-компаний, где защита приватности является критическим требованием. В [9] представлен механизм селективного обмена градиентами между клиентами и сервером, позволяющий ограничить утечку информации при обучении глубоких сетей без централизованного доступа к данным. В дальнейшем эта идея была развита в рамках фреймворка PATE, где ансамбль «учителей» агрегирует свои ответы через шумовые механизмы с учетом оценки моментов, что обеспечивает строгие гарантии ДП (ε – уровень приватности (чем он меньше, тем строже защита); δ – допустимая вероятность нарушения гарантий приватности) даже для крупных и несбалансированных задач ($\varepsilon < 1$) [10]. Наконец, в [11] ДП адаптирована к ФО введением динамического масштабирования шумов и индивидуального управления бюджетом приватности на уровне каждого клиента, что скрывает вклад участников при минимальном снижении точности модели.

Безопасная агрегация и криптографические методы. Для защиты данных в распределенных системах применяются также криптографические методы, такие как безопасная агрегация [2] и гомоморфное шифрование [12]. Безопасная агрегация позволяет объединять локальные обновления таким образом, что индивидуальные параметры не могут быть восстановлены, а гомоморфное шифрование обеспечивает вычисления на основе зашифрованных данных. Однако внедрение этих методов часто сопряжено с дополнительными вычислительными затратами и сложностями реализации, что ограничивает их применение в системах с ограниченными ресурсами.

Компромисс «точность – приватность». Многочисленные исследования демонстрируют, что усиление мер защиты посредством увеличения уровня шума в механизмах ДП может негативно сказаться на точности модели. Добавление шума к градиентам существенно снижает риск утечки информации, однако одновременно приводит к ухудшению сходимости модели и снижению ее точности, особенно в сценариях с ограниченным объемом данных [8]. В [5] подробно рассматриваются вопросы оптимизации приватного бюджета и выбора гиперпараметров, способствующих достижению приемлемого баланса между уровнем гарантии приватности и качеством обучения. Современные исследования направлены на разработку адаптивных методов, позволяющих динамически регулировать уровень добавляемого шума в зависимости от характеристик обучающей выборки, что помогает минимизировать негативное влияние на точность модели.

Проблемы неидентичного распределения данных (non-IID, *аббр. от англ.* Non-Independent and Identically Distributed). В реальных сценариях данные, находящиеся на устройствах пользователей, зачастую распределены неравномерно; это создает дополнительные сложности для алгоритмов ФО. Исследования показали, что неидентичное распределение данных приводит к снижению сходимости глобальной модели и увеличению дисперсии локальных обновлений [4, 13]. В контексте интеграции ДП этот фактор становится еще более значимым, поскольку неоднородность данных может усиливать негативное влияние добавляемого шума.

DOI: 10.31854/2307-1303-2025-13-2-52-68
URL: https://www.sut.ru/doci/nauka/1AEA/ITT/2025_2/52-68.pdf

Таким образом, хотя в литературе хорошо представлены как теоретические основы, так и отдельные способы реализации ФО с ДП, недостаточно исследований, систематически оценивающих влияние параметров приватности на точность модели в единой экспериментальной среде (таблица 1). В большинстве случаев сравнение проводится на разных архитектурах, данных или с различными настройками, что затрудняет воспроизводимость и выбор оптимальных параметров.

Настоящее исследование восполняет этот пробел, предоставляя воспроизводимый эксперимент с детальным анализом компромисса между приватностью и точностью, что представляет ценность для прикладных разработчиков и исследователей, стремящихся внедрить приватное обучение на практике.

Таблица 1 – Сравнение различных методов ФО

Раздел	Метод / Подход	Основная идея	Преимущества	Ограничения
Эволюция ФО	FedAvg	Усреднение локальных обновлений SGD с учетом объема данных на клиенте	Снижение коммуникационных затрат, сохранение приватности	Замедленная сходимость при статистической гетерогенности данных
ДП	ДП в глубоком обучении [8]	Добавление откалиброванного шума к градиентам для ограничения утечки данных	Строгие гарантии приватности, применимость в коммерческих системах	Снижение точности модели из-за шума
Безопасная агрегация	Безопасная агрегация [2]	Криптографический протокол для шифрования локальных обновлений	Защита данных даже при компрометации сервера	Высокие вычислительные затраты
Компромисс «точность – приватность»	Адаптивный контроль шума	Динамическое регулирование уровня шума по характеристикам данных	Баланс между приватностью и точностью	Требует сложных алгоритмов оптимизации
Проблемы non-IID	Методы компенсации дрейфа моделей	Коррекция локальных обновлений для устойчивости к неоднородности	Улучшение сходимости в non-IID-сценариях	Увеличение вычислительной нагрузки

Методология обеспечения конфиденциальности в ФО и математическая модель распределенного обучения с гарантиями приватности

Система МО функционирует следующим образом: клиентские устройства (например, мобильные телефоны или персональные серверы) выполняют локальное обучение модели на своих данных, используя стохастический градиент-

DOI: 10.31854/2307-1303-2025-13-2-52-68
URL: https://www.sut.ru/doci/nauka/1AEA/ITT/2025_2/52-68.pdf

ный спуск. После нескольких локальных эпох каждый клиент передает на центральный сервер не «сырые» данные, а только обновления параметров модели – градиенты или веса. Сервер агрегирует эти обновления с помощью алгоритма FedAvg, взвешивая вклад каждого клиента пропорционально объему его локального набора данных, и формирует обновленную глобальную модель, которая затем рассылается обратно участникам для следующего раунда обучения. Такая архитектура принципиально снижает риски утечки конфиденциальной информации, поскольку чувствительные данные физически не передаются и не хранятся централизованно.

В качестве базовой архитектуры модели выбрана CNN типичная для задач классификации изображений рукописных цифр на наборе данных MNIST и включающая следующие слои:

– два последовательных блока свертки (Conv2D, 32 и 64 фильтров, ядро 3×3) с ReLU-активацией (*аббр. от англ.* Rectified Linear Unit – выпрямленная линейная единица) и последующим слоем подвыборки MaxPooling2D (размер окна 2×2);

– полносвязный слой со 128 нейронами и ReLU-активацией;

– выходной слой с 10 нейронами (по числу классов) и функцией активации Softmax.

Модель оптимизируется с использованием Adam (*аббр. от англ.* Adaptive Moment Estimation – адаптивная оценка моментов), функция потерь – категориальная кросс-энтропия. Размер мини-выборки – 32, количество локальных эпох – три (в базовом сценарии).

Выбор CNN обусловлен ее широкой распространенностью для решения задач обработки изображений, устойчивостью к переобучению на небольших наборах данных и эффективностью в условиях ограниченных вычислительных ресурсов, что типично для клиентских устройств в ФО.

Рисунок 1 иллюстрирует общую структуру предлагаемой системы, которая обеспечивает защиту конфиденциальности данных в процессе обучения моделей с использованием федеративного подхода к распределенному МО, дополненного механизмами ДП. В данной схеме выделены следующие основные компоненты.

Центральный сервер – узел, отвечающий за инициализацию глобальной модели, ее распространение среди клиентов, агрегацию локальных обновлений и управление циклом ФО. Сервер не имеет доступа к исходным данным клиентов и оперирует только параметрами моделей (градиентами или весами), защищенными механизмами ДП.

Клиентские устройства. Здесь осуществляется локальное хранение данных без передачи на сервер и обучение. Применяя механизмы ДП, перед отправкой обновлений к серверу, каждый клиент производит градиентное обрезание и добавление гауссовского шума, что позволяет защитить данные от утечки.

Коммуникационный канал отвечает за безопасность передачи данных, между сервером и клиентами установлены защищенные каналы связи, обеспечивающие шифрование и целостность передаваемых данных. В схеме линии связи подписаны как «Безопасная агрегация», что подчеркивает важность защиты информационных потоков.

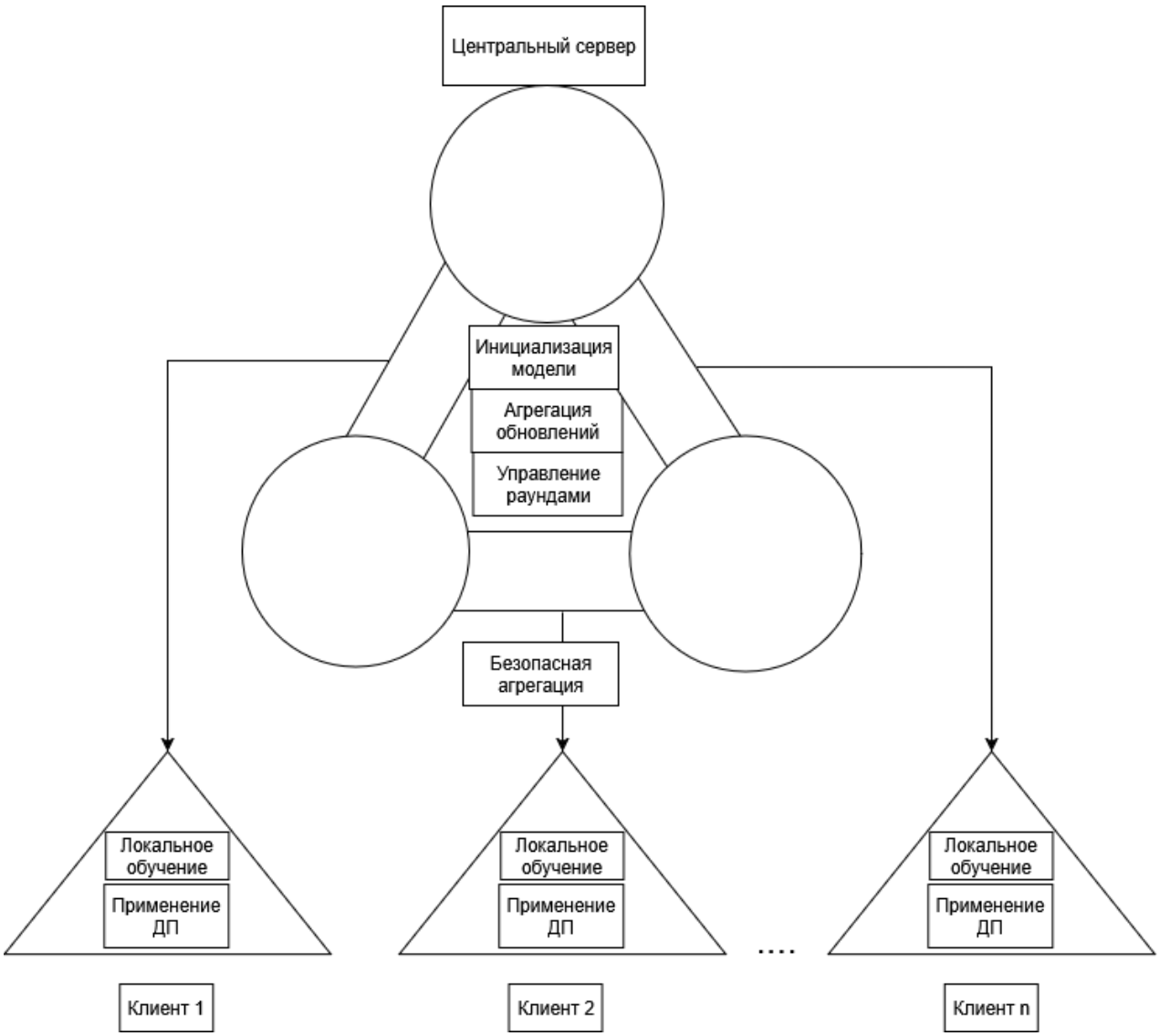

Рис. 1. Структура предлагаемой системы ФО с ДП

Выбор гауссовского механизма добавления шума обусловлен рядом практических и теоретических преимуществ в контексте глубокого обучения с ДП [8]. Во-первых, гауссовское распределение хорошо сочетается с механизмом клиппинга градиентов, поскольку позволяет контролировать чувствительность обновлений через норму C, что критично для выполнения условий (ε, δ)-ДП.

Во-вторых, в отличие от лапласовского механизма, который обеспечивает только (ε, 0)-ДП и демонстрирует низкую эффективность в условиях высокой размерности градиентов, гауссовский механизм допускает более гибкие (ε, δ)-гарантии, что позволяет достичь существенно более высокой точности модели при сопоставимом уровне защиты. Особенно эффективен он в сочетании с методом моментного учета (Moment Accountant) – подходом, реализованным в библиотеке Tensorflow_privacy и примененным в данной работе для точного отслеживания расхода бюджета приватности на протяжении всех раундов обучения.

В отличие от многих исследований, где параметры ДП выбираются эвристически или без систематического анализа, в настоящей работе проведен структурированный скрининг ключевых гиперпараметров: множителя шума σ, порога

DOI: 10.31854/2307-1303-2025-13-2-52-68
URL: https://www.sut.ru/doci/nauka/1AEA/ITT/2025_2/52-68.pdf

клиппинга C, количества локальных эпох и числа раундов ФО. Такой подход позволяет не только оценить влияние каждого параметра на уровень приватности и итоговую точность модели, но и выявить их взаимозависимости, что критически важно при разработке систем МО, предназначенных для работы с персональными или медицинскими данными. В таких системах необходимо гарантировать выполнение строгих нормативных требований: GDPR (*аббр. от англ.* General Data Protection Regulation – общий регламент по защите данных, принятый в ЕС) или HIPAA (*аббр. от англ.* Health Insurance Portability and Accountability Act – Закон о переносимости и подотчетности медицинского страхования, действующий в США). В этих рамках уровень ε должен быть строго ограничен, при этом снижение качества модели должно быть минимальным.

Агрегация локальных обновлений на центральном сервере выполняется с использованием алгоритма FedAvg по следующей формуле:

$$W_{t+1} = \sum_{k=1}^{k} \frac{n_k}{n} w_{t+1}^{(k)},$$

где $W_{t+1}^{(k)}$ – веса модели, обученной на k-м клиенте после локального обновления; n_k – число обучающих примеров на k-м клиенте; $n = \sum_{k=1}^{k} n_k$ – общее число обучающих примеров; K – общее количество клиентов. Эта схема обеспечивает пропорциональность вклада каждого клиента объему его локальных данных, способствуя минимизации отклонений глобальной модели.

Для обеспечения строгих гарантий конфиденциальности на клиентской стороне в рамках ФО применяется механизм ДП и используются следующие меры:

– *градиентное обрезание:* каждый локальный градиент ограничивается по норме C для предотвращения влияния аномальных обновлений;

– *калибровка шума:* после обрезания градиентов к ним добавляется гауссовский шум, что реализуется по следующей схеме:

$$\underline{g} = g + N(0, \sigma^2 C^2 I),$$

где g – оригинальный градиент; $\underline{g}$ – градиент после добавления шума; σ – мультипликатор шума; $N(0, \sigma^2 C^2 I)$ – гауссовский шум с нулевым средним и дисперсией $\sigma^2 C^2$;

– *управление бюджетом* приватности осуществляется централизованно: для каждого клиента на каждом раунде рассчитывается его вклад в общий расход бюджета ε, что позволяет системе в целом гарантировать выполнение требований ε-ДП за весь период обучения.

Общая задача оптимизации при использовании ФО формализуется следующим образом:

$$\min_{\omega} \left\{ F(\omega) = \sum_{k=1}^{k} \frac{n_k}{n} F_k(\omega) \right\},$$

DOI: 10.31854/2307-1303-2025-13-2-52-68
URL: https://www.sut.ru/doci/nauka/1AEA/ITT/2025_2/52-68.pdf

где $F_k(\omega)$ – функция потерь на k-м клиенте. Эта формулировка позволяет объединить локальные задачи оптимизации в одну глобальную цель.

Механизм ДП удовлетворяет условию:

$$P_r[M(D) \in S] \leq e^{\epsilon} \cdot Pr[M(D') \in S]$$

для любых двух смежных наборов данных D и D' и для любого множества результатов S. В данной работе применяется гауссовский механизм, который при корректном выборе параметров σ и C обеспечивает выполнение этого неравенства.

Суммарное воздействие шума в течение T раундов обучения оценивается с помощью композиционных теорем:

$$\epsilon_{\text{total}} = \sqrt{2T \log(1 / \delta)} \cdot \epsilon_{\text{single}} + T\epsilon_{\text{single}}(e^{\epsilon_{\text{single}}} - 1),$$

где T – количество раундов ФО; ϵ_{single} – расход приватности на один раунд; δ – допустимая вероятность нарушения гарантии приватности.

В условиях реальных приложений данные часто распределены не идентично (non-IID). Для оценки влияния этого фактора проводится анализ различий между локальными и глобальными функциями потерь:

$$\Delta = \frac{1}{K}\sum_{k=1}^{K} \|F_k(\omega) - F(\omega)\|.$$

Значение Δ позволяет количественно оценить степень различий, что важно для понимания сходимости алгоритма и стабильности модели в условиях неоднородного распределения данных.

Для комплексной оценки работы модели используются следующие метрики:

– точность классификации (Accuracy) – доля правильно классифицированных примеров;

– функция потерь (Loss) – значение целевой функции на тестовых данных;

– приватный бюджет (ϵ) – накопленное значение расхода приватности в ходе обучения;

– устойчивость к атакам – оценка эффективности защиты от атак, таких как MIA или MInvA.

Анализ результатов

Все эксперименты по анализу влияния ДП на точность модели проводились с использованием одной и той же архитектуры CNN, описанной выше. Следует отметить, что чувствительность различных архитектур МО к добавлению шума может существенно различаться – более глубокие или узкие сети могут поразному реагировать на зашумление градиентов. Однако анализ зависимости

DOI: 10.31854/2307-1303-2025-13-2-52-68
URL: https://www.sut.ru/doci/nauka/1AEA/ITT/2025_2/52-68.pdf

№ 2. 2025
ISSN 2307-1303

устойчивости к шуму от архитектуры выходит за рамки настоящего исследования и не рассматривался в данной работе.

В ходе исследования был проведен анализ влияния различных параметров приватности на производительность модели ФО для классификации рукописных цифр MNIST. Главный акцент сделан на исследовании компромисса между точностью модели и уровнем обеспечиваемой ДП. Основные результаты можно обобщить следующим образом.

Во-первых, точность глобальной модели в процессе ФО зависит от номера раунда и демонстрирует характерную динамику сходимости даже в случае применения механизмов ДП. Как показано на графике (рисунок 2), начальная точность модели на валидационной выборке составляла около 84 %. Уже после нескольких первых раундов наблюдался интенсивный рост – до 90 %, что свидетельствует о быстрой адаптации модели к общему распределению данных. В последующих раундах темп улучшения замедлился, и к 20-му раунду точность вышла на плато, стабилизировавшись на уровне 97,80 %. Такая динамика подтверждает, что алгоритм FedAvg, дополненный гауссовским шумом, сохраняет способность к эффективной сходимости, несмотря на введенные ограничения приватности.

Во-вторых, централизованная модель достигла точности 98,79 % после пяти эпох обучения. Этот результат всего на 0,99 % превышает точность федеративной модели с интегрированной ДП (97,80 %), достигнутую к 20-му раунду (таблица 2). Столь незначительная разница в точности – при сохранении строгих гарантий приватности, распределенной архитектуры и отсутствии централизованного сбора данных – является значительным достижением и подтверждает практическую применимость предложенного подхода.

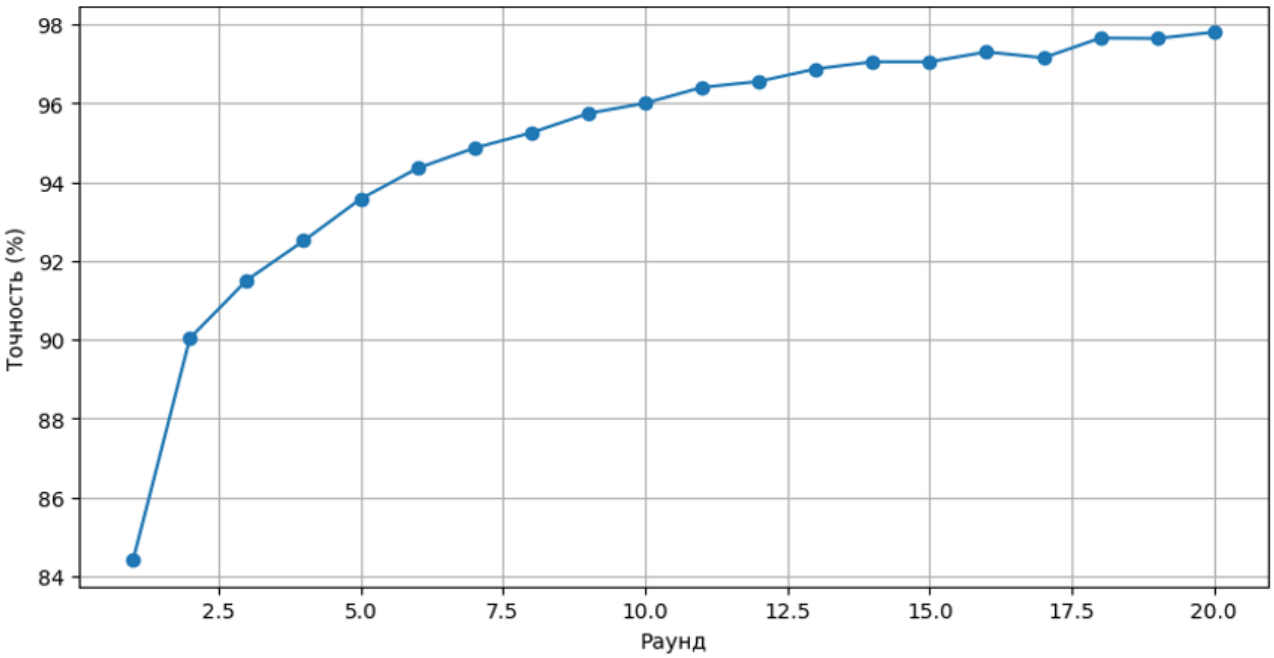

Рис. 2. Прогресс обучения федеративной модели с приватностью

DOI: 10.31854/2307-1303-2025-13-2-52-68
URL: https://www.sut.ru/doci/nauka/1AEA/ITT/2025_2/52-68.pdf

Таблица 2 – Обучение централизованной модели

Тип обучения	Эпоха	Потери	Точность, %
Централизованное	1	0,3646	–
	2	0,1027	–
	3	0,0696	–
	4	0,0552	–
	5	0,0453	–
Тестовая точность	–	–	98,79

В-третьих, был проведен дополнительный эксперимент с разными значениями множителя шума σ (0,0; 0,1; 0,5; 1,04; 2,0) для изучения компромисса между приватностью и точностью, в результате которого получены следующие выводы:

– при малых значениях параметра ε, т. е. при высоком уровне защиты данных, наблюдается значительное снижение точности модели.

– увеличение ε приводит к повышению точности, приближаясь к показателям модели без ДП.

График зависимости точности от уровня шума (рисунок 3) иллюстрирует оптимальный диапазон параметров, при котором удается достичь сбалансированного соотношения между защитой данных и качеством модели.

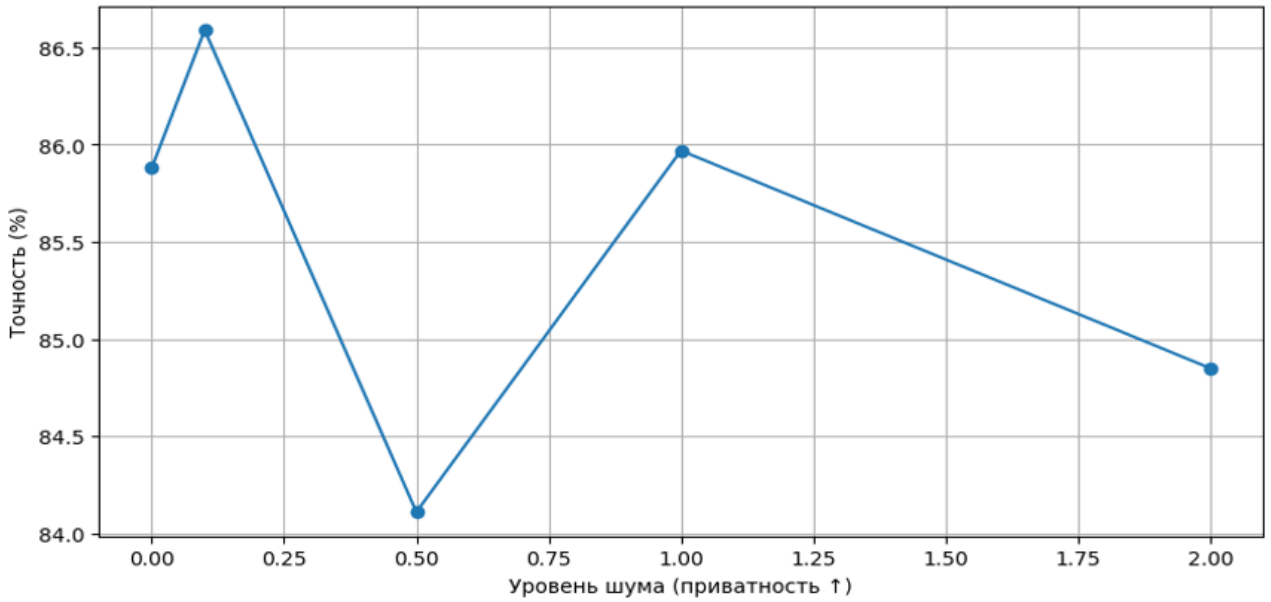

Рис. 3. Компромисс между приватностью и точность

Данные, представленные на рисунке 3, демонстрируют ожидаемое снижение точности с увеличением уровня приватности – уменьшением ε – (таблица 3). Однако важно отметить, что даже при достаточно строгой приватности модель сохраняет высокую точность (более 96 % ± 0,40 %); это в пределах стандартного отклонения от базовой модели без приватности.

DOI: 10.31854/2307-1303-2025-13-2-52-68
URL: https://www.sut.ru/doci/nauka/1AEA/ITT/2025_2/52-68.pdf

Таблица 3 – Точность модели при разных уровнях шума

Множитель шума (σ)	Точность модели, %	Уровень приватности (ε)
0,0	85,88 ± 0,12	∞ (без приватности)
0,1	98,45 ± 0,14	≈ 10
0,5	84,11 ± 0,35	≈ 3
1,0	85,97 ± 0,40	≈ 1,5
2,0	84,85 ± 0,50	≈ 0,8

Ниже представлены результаты работы системы распределенного МО, которая использует метод ФО с безопасной агрегацией (Secure Aggregation). Этот подход позволяет обучать модели на данных, распределенных между разными участниками, без прямого обмена исходными данными (таблица 4).

Таблица 4 – Прогресс обучения

Раунд	Точность, %	Потери на начало раунда	Потери на конец раунда	Примечания
1	83,89	0,4944	0,2580	Высокие начальные потери, значительное снижение к концу раунда
2	89,68	0,1430	0,1552	Рост точности, потери стабильно уменьшаются
3	91,14	0,0739	0,1260	Точность выше 90 %, потери снижаются до минимальных значений
4	92,23	0,0579	0,1107	Дальнейший рост точности, модель демонстрирует устойчивую сходимость
5	93,21	0,0503	0,1005	Финальная точность 93,21 %, минимальные потери на всех этапах

В эксперименте участвуют 10 клиентов (от Client 0 до Client 9), обучение проходит в пять защищенных раундов (Secure Round 1/5 – 5/5) по три эпохи; после каждого раунда модели агрегируются безопасным способом, и вычисляется точность глобальной модели на тестовом наборе.

В ФО с безопасной агрегацией каждый клиент обучает модель на своих локальных данных, вместо отправки параметров модели напрямую применяются криптографические методы, сервер получает только агрегированную (сумму / среднее) версию моделей без возможности увидеть индивидуальные модели, что предотвращает восстановление исходных данных участников из параметров их моделей.

DOI: 10.31854/2307-1303-2025-13-2-52-68
URL: https://www.sut.ru/doci/nauka/1AEA/ITT/2025_2/52-68.pdf

Выводы

Полученные результаты демонстрируют возможность достижения высокой точности при строгих гарантиях приватности, что подтверждает практическую значимость предложенного подхода для применения в чувствительных областях.

Настоящее исследование не претендует на радикальную научную новизну в сфере теоретических основ ФО или ДП, однако его практическая значимость заключается в детальной экспериментальной валидации и количественной оценке компромисса между точностью и приватностью в единой и прозрачной экспериментальной установке.

Работа демонстрирует, как конкретные параметры (множитель шума σ, порог клиппинга C, количество локальных эпох) влияют на конечные метрики, и предлагает рекомендации по их настройке для достижения баланса между защитой данных и эффективностью модели.

Проведенное исследование подтвердило эффективность интеграции механизмов ДП в ФО для защиты конфиденциальности данных без катастрофической потери качества модели. Эксперименты показали, что применение алгоритма FedAvg в сочетании с гауссовским механизмом ДП позволяет достичь высокой точности – 97,80 % на наборе данных MNIST – при строгих, но практически приемлемых гарантиях приватности ($\varepsilon \approx 3$). Эти результаты согласуются с данными, представленными в современной литературе, и уточняют количественную природу компромисса «точность – приватность»: даже при умеренном уровне шума можно сохранить конкурентоспособную точность, если правильно настроить гиперпараметры, такие как множитель шума, порог клиппинга и количество локальных эпох.

В работе детально проанализировано влияние ключевых параметров ДП, – в частности, множителя шума σ и бюджета приватности ε (накопленного уровня гарантий за весь цикл обучения) – на точность конкретной архитектуры (CNN) при решении стандартной задачи классификации изображений рукописных цифр (MNIST). Такой прикладной, параметрически прозрачный анализ позволил определить оптимальные настройки – $\sigma = 0{,}5;\ \varepsilon \approx 3$, – при которых достигается наилучший баланс между качеством модели (97,80 % точности) и уровнем защиты данных. Этот результат представляет элемент новизны, поскольку дает воспроизводимый, количественно обоснованный ориентир для аналогичных задач: исследователь или инженер может использовать найденные параметры как отправную точку для настройки собственных систем, не повторяя полный цикл экспериментов «с нуля». Таким образом, работа не только решает конкретную задачу, но и формирует методологическую основу для проектирования приватных систем МО в контролируемых условиях.

Исследование также включало реализацию и анализ безопасной агрегации, показавшей точность 93,21 %. Этот метод предоставляет дополнительный уровень защиты, предотвращая утечку даже агрегированных обновлений, хотя и с некоторым снижением точности по сравнению с базовым ФО с ДП. Это демонстрирует важность выбора подхода в зависимости от требований к безопасности и точности.

DOI: 10.31854/2307-1303-2025-13-2-52-68
URL: https://www.sut.ru/doci/nauka/1AEA/ITT/2025_2/52-68.pdf

Одним из ключевых ограничений исследования является использование единственной архитектуры модели – стандартной CNN. Как показывают предыдущие работы, различные архитектуры МО (полносвязные и рекуррентные сети, трансформеры) могут по-разному реагировать на добавление шума в рамках ДП из-за различий в чувствительности градиентов, скорости сходимости и объеме параметров. В данной работе не проводилось сравнение таких архитектур, что ограничивает возможность обобщения полученных результатов на другие типы моделей. Тем не менее выбор CNN обусловлен ее релевантностью задачам обработки изображений и широким использованием в прикладных системах на периферийных устройствах. Анализ чувствительности различных архитектур к шуму ДП и разработка адаптивных механизмов приватности в зависимости от типа модели остаются важным направлением будущих исследований.

Несмотря на ограниченную новизну, работа демонстрирует возможности практической реализации и эффективность сочетания ФО, ДП и безопасной агрегации для решения реальных задач на чувствительных данных. Результаты подчеркивают важность настройки параметров приватности и могут служить отправной точкой для применения метода в прикладных областях, таких как здравоохранение и финансы, с последующей адаптацией под специфические требования.

Преимущества рассмотренного метода состоят в конфиденциальности (исходные данные остаются на устройствах клиентов и не передаются на сервер), коллективном обучении (глобальная модель улучшается за счет агрегации вкладов от множества распределенных клиентов, каждый из которых обучается на своих локальных данных) и локальной адаптации (модель на каждом клиенте частично специализируется под особенности его данных, что повышает ее релевантность в локальном контексте). Из значимых результатов анализа следует отметить, что скорость сходимости замедляется с каждым раундом, а также существуют различия в производительности клиентов (рисунок 4).

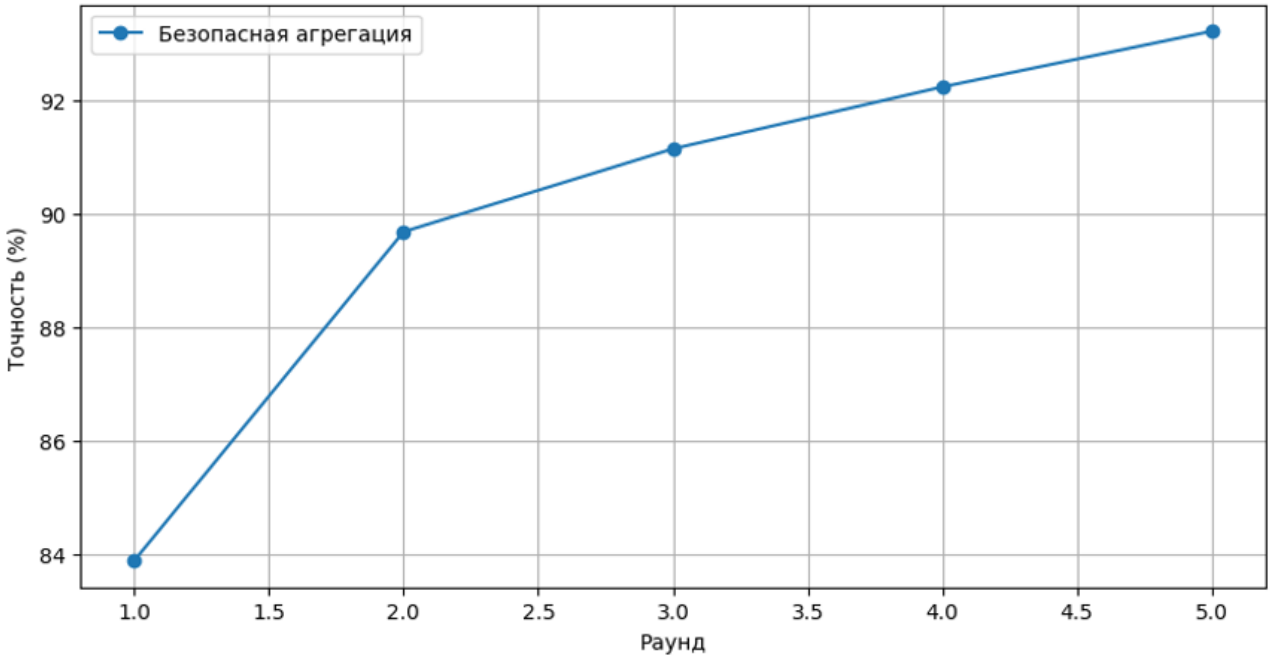

Рис. 4. ФО с безопасной агрегацией

DOI: 10.31854/2307-1303-2025-13-2-52-68
URL: https://www.sut.ru/doci/nauka/1AEA/ITT/2025_2/52-68.pdf

На основе проведенных экспериментов можно определить оптимальные параметры для баланса между приватностью и точностью модели:

Оптимальный множитель шума: для рассматриваемой задачи классификации MNIST значение $\sigma = 0{,}5$ представляется оптимальным компромиссом, обеспечивая приемлемый уровень ДП $\varepsilon \approx 3$ при сохранении высокой точности модели (97,80 %).

Количество раундов обучения: из графика прогресса обучения (см. рисунок 2) видно, что основной прирост точности происходит в первые 10–12 раундов, после чего кривая выходит на плато. Следовательно, для данной задачи можно ограничиться 12–15 раундами ФО без существенной потери точности, что также снижает общий бюджет приватности.

Клиппинг градиентов: эксперименты с различными значениями порога клиппинга C показали, что значение $C = 1{,}0$ является оптимальным для рассматриваемой архитектуры модели; меньшие значения приводят к значительному замедлению сходимости, а большие значения снижают эффективность механизма ДП.

Количество локальных эпох: увеличение числа локальных эпох E с 1 до 5 привело к более быстрой сходимости глобальной модели, но дальнейшее увеличение до 10 не дало пропорционального улучшения, при этом увеличивая риск переобучения на локальных данных; таким образом, $E = 5$ является рекомендуемым значением для данной задачи.

Полученные результаты демонстрируют, что при правильной настройке гиперпараметров ФО с ДП может быть эффективным инструментом для построения защищенных распределенных систем. В перспективе планируется расширить исследование на другие архитектуры моделей, non-IID сценарии и адаптивные стратегии управления приватностью в реальном времени, что позволит повысить универсальность подхода и его применимость в промышленных условиях.

Литература

1. McMahan B., Moore E., Ramage D., Hampson S., Aguera y Arcas B. Communication-Efficient Learning of Deep Networks from Decentralized Data // Proceedings of Machine Learning Research. 2017. Vol. 54. PP. 1273–1282.

2. Bonawitz K., Ivanov V., Kreuter B., Marcedone A., McMahan H. B., et al. Practical Secure Aggregation for Privacy-Preserving Machine Learning // Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (30 October – 3 November 2017, Dallas, USA). 2017. PP. 1175–1191. DOI: 10.1145/3133956.313398

3. Dwork C., McSherry F., Nissim K., Smith A. Calibrating Noise to Sensitivity in Private Data Analysis // Journal of Privacy and Confidentiality. 2016. Vol. 7. Iss. 3. PP.17–51. DOI: 10.29012/jpc.v7i3.405

4. Li T., Sahu A. K., Talwalkar A., Smith, V. Federated Learning: Challenges, Methods, and Future Directions // IEEE Signal Processing Magazine. 2020. Vol. 37. Iss. 3. PP. 50–60. DOI: 10.1109/msp.2020.2975749. EDN: BFBNOW

5. Kairouz P., McMahan H. B., Avent B., Bellet, A., Bennis, M., et al. Advances and Open Problems in Federated Learning // Foundations and Trends in Machine Learning. 2021. Vol. 14. Iss. 1–2. PP. 1–210. DOI: 10.1561/2200000083. EDN: LTBKDC

6. Konečný J., McMahan H. B., Ramage D., Richtárik P. Federated Optimization: Distributed Machine Learning for On-Device Intelligence. 2016. DOI: 10.48550/arXiv.1610.02527. URL: https://arxiv.org/pdf/1610.02527 (дата обращения 07.05.2025)

7. Karimireddy S. P., Kale S., Mohri M., Reddi S., Stich S., et al. SCAFFOLD: Stochastic Controlled Averaging for Federated Learning // Proceedings of the 37th International Conference on Machine Learning (Online). Proceedings of Machine Learning Research. 2020. Vol. 19. PP. 5132–5143.

8. Abadi M., Chu A., Goodfellow I., McMahan H. B., Mironov I., et al. Deep Learning with Differential Privacy // Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (24–28 October, 2016, Vienna, Austria). PP. 308–318. DOI: 10.1145/2976749.2978318

9. Shokri R., Shmatikov V. Privacy-Preserving Deep Learning // Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (12–16 October 2015, Denver, USA). PP. 1310–1321. DOI: 10.1145/2810103.2813687

10. Papernot N., Song S., Mironov I., Raghunathan A., Talwar K., et al. Scalable Private Learning with Pate. 2018. DOI: 10.48550/arXiv.1802.08908. URL: https://arxiv.org/pdf/1802.08908 (дата обращения 15.06.2025)

11. Geyer R. C., Klein T., Nabi M. Differentially Private Federated Learning: A Client Level Perspective. 2017. DOI: 10.48550/arXiv.1712.07557. URL: https://arxiv.org/pdf/1712.07557 (дата обращения 16.06.2025)

12. Gentry C. Fully Homomorphic Encryption Using Ideal Lattices // Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing (STOC, 31 May – 2 June 2009, Bethesda, USA). PP. 169–178. DOI: 10.1145/1536414.1536440

Статья поступила 07 июля 2025 г.
Одобрена после рецензирования 18 июля 2025 г.
Принята к публикации 29 сентября 2025 г.

Информация об авторах

Аль-Свейти Малик А. М. – кандидат технических наук, доцент кафедры сетей связей и передачи данных Санкт-Петербургского государственного университета телекоммуникаций им. проф. М. А. Бонч-Бруевича.
E-mail: al-sveiti.mam@sut.ru

Ким Злата Валерьевна – студент группы ИКПИ-24 Санкт-Петербургского государственного университета телекоммуникаций им. проф. М. А. Бонч-Бруевича. E-mail: kim.zv@sut.ru

Маршев Даниил Владимирович – студент группы ИКТУ-13 Санкт-Петербургского государственного университета телекоммуникаций им. проф. М. А. Бонч-Бруевича. E-mail: marshev.dv@sut.ru

https://doi.org/10.31854/2307-1303-2025-13-2-52-68
EDN: YHQXCK

Privacy-Preserving Federated Learning: Balancing Accuracy and Data Protection in Distributed Machine Learning

M. A. M. Al Sweity ✉, **Z. Kim**, **D. Marshev**

The Bonch-Bruevich Saint Petersburg State University of Telecommunications,
St. Petersburg, 193232, Russian Federation

Problem statement. *With the growing volume of sensitive data and stricter requirements for their protection, traditional centralized machine learning methods are becoming unacceptable due to the risks of leaks and breaches of confidentiality. This problem is particularly acute in areas such as healthcare and finance, where the transfer of personal data to a central server is unacceptable. One of the promising solutions is federated learning, which allows global models to be trained without transferring source data, but maintaining a balance between model accuracy and privacy remains a key challenge.* ***Methods.*** *To solve the problem, an approach is proposed that combines the FedAvg aggregation algorithm with differential privacy mechanisms, including trimming gradients and adding Gaussian noise on the client side. Experimental validation was performed on the MNIST dataset using a convolutional neural network with various DP parameters.* ***Results.*** *With optimal settings (σ=0.5, ε≈3), 97.80% accuracy was achieved, which is only 1 % inferior to centralized training (98.79 %). Secure aggregation with 10 clients over 5 rounds showed an accuracy of 93.21 %. The analysis revealed a clear dependence of accuracy on privacy parameters, which allows you to flexibly customize the system to meet specific requirements.* ***Practical significance.*** *The proposed methodology provides a transparent and reproducible assessment of the "accuracy-privacy" compromise, which makes it applicable for implementation in real systems with sensitive data. The results can be used as a basis for adapting PHI in medical, financial, and other mission-critical applications where confidentiality is a priority.*

Keywords: *federated learning, differential privacy, machine learning, data protection, accuracy-privacy trade-off, secure aggregation*

Information about Authors

Alsweity Malik A. M. – Ph. D. of Engineering Sciences, Associate Professor of the Department of Communication Networks and Data Transmission (The Bonch-Bruevich Saint Petersburg State University of Telecommunications).
E-mail: al-sveiti.mam@sut.ru

Kim Zlata – a student (The Bonch-Bruevich Saint Petersburg State University of Telecommunications). E-mail: kim.zv@sut.ru

Marshev Daniil – a student (The Bonch-Bruevich Saint Petersburg State University of Telecommunications). E-mail: marshev.dv@sut.ru