**УДК 004.89**

# СРАВНЕНИЕ МЕТОДОВ ДИФФЕРЕНЦИАЛЬНОЙ ПРИВАТНОСТИ В ЗАДАЧАХ ДОВЕРЕННОГО ОБУЧЕНИЯ

Ситдиков Д.С., младший научный сотрудник
Военная академия связи имени Маршала Советского Союза С. М. Буденного,
г. Санкт-Петербург

Дифференциальная приватность (DP) на сегодняшний день является одним из наиболее фундаментальных и формально обоснованных подходов к обеспечению конфиденциальности данных в системах машинного обучения [1].

В условиях, когда всё большее число моделей обучаются на чувствительной информации – медицинских записях, пользовательском поведении, финансовых транзакциях и т.п. – потребность в надёжных механизмах защиты становится не только технической, но и юридической и этической обязанностью.

В этой связи задачи доверенного обучения всё чаще требуют применения дифференциальной приватности как средства, гарантирующего, что участие отдельного пользователя в обучающей выборке не оказывает значимого влияния на поведение модели и, следовательно, не может быть выявлено или скомпрометировано внешними наблюдателями или злоумышленниками [2].

Сравнение методов дифференциальной приватности в контексте доверенного обучения представляет собой исследовательскую задачу, охватывающую различные аспекты: архитектурные особенности моделей, алгоритмы оптимизации, параметры шума, механизмы отслеживания приватных бюджетов, а также баланс между уровнем приватности и точностью модели. Главные различия между методами дифференциальной приватности обусловлены тем, на каком этапе и с какой точностью они внедряются: на уровне входных данных, градиентов, весов модели, выходов или целого обучения в целом.

Одним из наиболее распространённых и практически применяемых подходов является дифференциально-приватный стохастический градиентный спуск (DP-SGD). Суть его состоит в том, что при каждом обновлении модели градиенты, рассчитанные на мини-пакетах данных, сначала обрезаются (clipping) до заданной нормы (L2-нормы), а затем к ним добавляется гауссов шум с параметрами, зависящими от желаемого уровня приватности. DP-SGD позволяет контролировать параметры ε (эпсилон) и δ (дельта), которые задают математическую степень конфиденциальности, и при правильной настройке обеспечивает строгие границы на влияние отдельных точек данных. Однако

DP-SGD имеет ряд недостатков: он чувствителен к размеру батча, требует большого числа итераций для достижения приемлемой точности, а шум может сильно снижать качество модели, особенно на малых выборках. Сравнительные эксперименты показывают, что эффективность DP-SGD зависит от архитектуры: простые линейные модели устойчивее к шуму, тогда как глубокие нейронные сети демонстрируют существенное падение точности при малых значениях ε.

Другим подходом является дифференциально-приватное обучение с использованием механизмов постобработки. Здесь предполагается, что приватность обеспечивается не в процессе обучения, а при предоставлении выходов модели, либо в процессе публикации параметров. Например, можно публиковать только агрегированные модели, усреднённые по участникам (в случае federated learning), либо добавлять шум к самим параметрам модели перед развёртыванием. Преимущество такого подхода – минимальное влияние на сам процесс обучения и возможность сохранения высокой точности. Однако это требует более сложных теоретических обоснований, поскольку гарантия приватности обеспечивается лишь на конечном этапе, и утечки могут произойти в процессе самого обучения, если используется неподготовленная инфраструктура [3].

Также существует метод приватных агрегатов (Private Aggregation of Teacher Ensembles – PATE), основанный на идее использования нескольких предварительно обученных моделей-учителей, которые обучаются на независимых разбиениях приватных данных. Их ответы по меткам затем агрегируются с добавлением шума, чтобы обучить модель-ученика. Такой метод обеспечивает сильные теоретические гарантии приватности и устойчив к запоминанию конкретных примеров, но плохо масштабируется на большие данные и требует существенных вычислительных ресурсов. В задачах с ограниченным числом классов и сравнительно компактными архитектурами он показывает высокую эффективность, однако в крупных NLP-моделях или визуальных задачах его применение затруднено.

Сравнение этих методов требует использования единой системы метрик. Основной метрикой приватности остаётся параметр ε, однако его интерпретация требует контекста: в одних приложениях допустимым может считаться $\varepsilon=5$, в других – только $\varepsilon<1$. Одновременно сравниваются точность модели, скорость обучения, стабильность градиентов, устойчивость к атакам восстановления данных (например, membership inference). Эксперименты показывают, что при использовании сильных регуляризаторов, увеличении размеров батчей и адаптивной нормализации градиентов (например, через оптимизаторы типа DP-Adam), DP-SGD может демонстрировать приемлемое соотношение точности и

приватности даже на сложных задачах. Однако с ростом глубины моделей или дисбаланса в данных эффективность снижается.

Отдельное направление – сравнение методов дифференциальной приватности в распределённом обучении, особенно в системах federated learning. Здесь модели обучаются на устройствах пользователей, а сервер агрегирует обновления, не получая сами данные. В такой архитектуре приватность может нарушаться через градиенты, поэтому DP используется в агрегации и/или на клиентской стороне. Методы с централизованным шумом показывают лучшую сходимость, но требуют доверия к серверу; локальные механизмы более устойчивы, но вызывают деградацию качества [4].

В заключение, выбор метода дифференциальной приватности в доверенном обучении зависит от множества факторов: размера и чувствительности данных, типа модели, наличия вычислительных ресурсов, правовых требований и области применения. Универсального решения пока не существует: каждая архитектура требует адаптации, настройки параметров и анализа последствий внедрения DP.

Однако общее направление однозначно указывает на рост интереса к формализуемым, воспроизводимым и интерпретируемым методам защиты, которые могут быть стандартизированы и использованы в рамках нормативных требований. Сравнительный анализ таких методов даёт разработчикам ценный инструмент оценки компромиссов между конфиденциальностью, точностью и устойчивостью моделей, тем самым формируя основу для по-настоящему доверенного машинного обучения [5].

**Список литературы:**

1. Rogers, R. M. Privacy Odometers and Filters: Pay-as-You-Go Composition / R. M. Rogers, A. Roth, J. Ullman, S. P. Vadhan // Advances in Neural Information Processing Systems. – 2016. – Vol. 29. – P. 1921–1929.

2. Dong, J. Gaussian Differential Privacy / J. Dong, A. Roth, W. J. Su // Journal of the Royal Statistical Society: Series B. – 2022. – Vol. 84, № 1. – P. 3–37. – DOI 10.1111/rssb.12454.

3. Balle, B. Improving the Gaussian Mechanism for Differential Privacy: Analytical Calibration and Optimal Denoising / B. Balle, Y.-X. Wang // Proceedings of the 35th International Conference on Machine Learning. – 2018. – Vol. 80. – P. 394–403.

4. Wang, Y.-X. Subsampled Rényi Differential Privacy and Analytical Moments Accountant / Y.-X. Wang, B. Balle, S. P. Kasiviswanathan // Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics. – 2019. – Vol. 89. – P. 1226–1235.

5. Papernot, N. Semi-Supervised Knowledge Transfer for Deep Learning from Private Training Data / N. Papernot, M. Abadi, Ú. Erlingsson [et al.] // International Conference on Learning Representations. – 2017. – P. 1–12.