УДК 004.89

АНАЛИЗ ГРАНИЦ ПРИВАТНОСТИ ПРИ МНОГОКРАТНОМ ИСПОЛЬЗОВАНИИ МОДЕЛИ В УСЛОВИЯХ ПЕРЕКРЫВАЮЩИХСЯ ДАННЫХ

Ким Д., студент направления Computer Visions Master's Degree Programm, II курс
Открытый университет Каталонии, г. Барселона, Испания

Повторное использование предобученных моделей в различных задачах становится основой современной инженерии машинного обучения. Такая практика оправдана как с точки зрения затрат на обучение, так и с позиций скорости внедрения решений, особенно при использовании трансформеров, больших языковых моделей и других ресурсоемких архитектур. Однако в контексте приватности данных это повторное использование порождает новые классы угроз, особенно если данные разных задач частично перекрываются или представляют одних и тех же пользователей в различных контекстах [1].

Анализ границ приватности в условиях многократного применения моделей на перекрывающихся данных становится ключевым элементом построения доверенных ИИ-систем, особенно в задачах, связанных с медициной, юриспруденцией, финансовыми сервисами и государственной аналитикой.

Границы приватности в данной постановке следует понимать как предел, до которого можно использовать модель, не допуская утечки информации о предыдущих задачах или обучающих примерах. Эта граница зависит как от внутренней структуры модели, так и от способа её настройки и степени перекрытия между наборами данных. В частности, повторное использование модели на частично совпадающих выборках может приводить к утечкам, даже если напрямую одни и те же данные не передаются между задачами. Это связано с тем, что нейросеть – особенно если она обладает высокой емкостью и была обучена без методов дифференциальной приватности – может сохранять параметры, чувствительные к уникальным образцам или распределениям.

Особенно уязвимыми становятся сценарии, в которых одни и те же субъекты данных встречаются в разных доменах. Например, пациент может быть представлен в медицинской задаче по распознаванию изображений, в другой – по анализу текстов жалоб, в третьей – по поведенческой модели использования мобильного приложения. Даже при отсутствии явной идентификации пользователь может быть «узнан» моделью по косвенным признакам. Это означает, что в случае повторной адаптации модели, получившей доступ к

дополнительной информации о пользователе, существует риск непреднамеренного объединения этих сведений и генерации нового знания, нарушающего границы приватности.

Для выявления и анализа таких ситуаций используются как теоретические, так и эмпирические методы [2].

Теоретические основы базируются на оценке взаимной информации между параметрами модели и чувствительными входами, на построении моделей угроз и сценариев атак, а также на оценке дивергенции распределений скрытых представлений. Особенно важным показателем служит способность модели выделять идентифицирующие признаки в случае даже минимального перекрытия данных между задачами. Такая способность может быть протестирована через атаки на приватность, включая membership inference (определение, принадлежал ли данный пример обучающей выборке) и model inversion (восстановление характерных входов по выходам модели).

Эмпирическая часть анализа включает создание тестов, имитирующих перекрытие данных различной плотности, с изменением степени перекрытия в контролируемых пределах.

Эти тесты позволяют зафиксировать критические пороги – уровни совпадения данных, при которых начинается утечка информации. Отдельное внимание уделяется изучению поведения модели в точках резкого ухудшения показателей приватности при незначительном изменении структуры входных данных. Такие пороги считаются индикаторами «границ приватности» и могут быть включены в протоколы проверки моделей перед повторным использованием [3].

Дополнительный риск возникает при использовании методов дообучения (fine-tuning) на новых задачах. Если новая выборка не полностью независима, а содержит пересечения по субъектам, то градиенты могут усилить закодированные следы прошлых обучающих данных. Это особенно критично в сценариях transfer learning, где сохраняются нижние слои модели, а верхние адаптируются под новые цели.

В таких случаях «наследуемые» параметры могут неосознанно переносить приватные паттерны, что приводит к новым каналам утечки. Некоторые исследования показывают, что даже агрегированные эмбеддинги могут содержать достаточно информации, чтобы восстановить детали оригинальных примеров в условиях перекрытия.

В контексте борьбы с такими угрозами разрабатываются методы динамической оценки границ приватности и адаптивного регулирования степени доступа к модели. Один из подходов включает в себя мониторинг влияния новых данных на внутренние представления сети: если активации скрытых слоёв демонстрируют избыточную чувствительность к частично повторяющимся

шаблонам, это может сигнализировать о потенциальной угрозе приватности. Другой подход связан с внедрением процедур «забвения» (machine unlearning), когда модель корректируется таким образом, чтобы минимизировать влияние уже обученных, но теперь запрещенных данных [4].

Неотъемлемым аспектом становится также нормативная сторона анализа границ приватности. Современные стандарты, включая GDPR, CCPA и локальные регламенты, требуют возможности удаления и отчуждения персональных данных, в том числе в том виде, в каком они могли быть сохранены в параметрах модели. Повторное использование моделей в условиях перекрытия может нарушить это требование, если отсутствуют механизмы контроля зависимости между новыми задачами и предыдущими обучающими выборками.

Таким образом, формализация границ приватности – не только техническая, но и правовая задача.

В заключение, можно утверждать, что в эпоху повторного и масштабируемого применения моделей ИИ границы приватности становятся динамичным, контекстно зависимым понятием. Их выявление, формализация и мониторинг требуют нового класса инструментов и протоколов, способных обеспечивать безопасность и доверие без снижения эффективности решений.

Этот вызов требует междисциплинарного подхода, сочетающего техническую строгость, юридическую осведомленность и этическую ответственность [5].

Список литературы:

1. Humbert, M. Quantifying Interdependent Risks in Genomic Privacy / M. Humbert, K. Huguenin, J. Hugonot [et al.] // ACM Transactions on Privacy and Security. – 2017. – Vol. 20, № 1. – Article 3. – DOI 10.1145/3035537.

2. Shomorony, I. Information-Theoretic Privacy in Genomic Databases / I. Shomorony, A. S. Avestimehr // IEEE International Symposium on Information Theory. – Piscataway : IEEE, 2015. – P. 1432–1436. – DOI 10.1109/ISIT.2015.7282670.

3. Fredrikson, M. Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing / M. Fredrikson, E. Lantz, S. Jha [et al.] // Proceedings of the 23rd USENIX Security Symposium. – Berkeley : USENIX Association, 2014. – P. 17–32.

4. Song, C. Machine Learning Models that Remember Too Much / C. Song, T. Ristenpart, V. Shmatikov // Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. – New York : ACM, 2017. – P. 587–601. – DOI 10.1145/3133956.3134077.

5. Jayaraman, B. Evaluating Differentially Private Machine Learning in Practice / B. Jayaraman, D. Evans // Proceedings of the 28th USENIX Security Symposium. – Berkeley : USENIX Association, 2019. – P. 1895–1912.