УДК 004.89

АНАЛИЗ КОМПРОМЕТАЦИИ КОНФИДЕНЦИАЛЬНЫХ ДАННЫХ ЧЕРЕЗ ОБРАТНЫЙ ИНЖИНИРИНГ МОДЕЛИ

Нгюэн З., студент направления BSc Information and Communication Technology, II курс
Королевский технологический институт, г. Стокгольм, Швеция

В современных системах искусственного интеллекта, основанных на глубоких нейронных сетях и других сложных моделях машинного обучения, обеспечение конфиденциальности данных является одной из наиболее критичных задач [1].

В частности, анализ компрометации конфиденциальной информации через методы обратного инжиниринга моделей приобретает особое значение на фоне растущей сложности моделей и увеличения объёмов обрабатываемых данных.

Обратный инжиниринг модели – это процесс, в ходе которого атакующий, обладая доступом к модели (например, к её параметрам, выходным данным или API), пытается восстановить исходные данные или выявить чувствительные паттерны, использованные при обучении. Исследование таких атак и методов защиты от них становится необходимым элементом создания доверенных и безопасных ИИ-систем [2].

Обратный инжиниринг моделей включает в себя широкий спектр техник, направленных на извлечение информации о структуре, параметрах и поведении модели с целью получения несанкционированного доступа к исходным данным или конфиденциальной информации. Классические методы включают model inversion attacks, при которых атакующий восстанавливает представления входных данных, используя выходы модели, а также membership inference attacks, позволяющие определить, присутствовал ли конкретный пример в обучающей выборке.

Современные подходы расширяются и усложняются, включая использование генеративных моделей для аппроксимации поведения целевой модели, методы активного обучения для выявления уязвимостей, а также комбинированные стратегии с применением вспомогательных данных и метаинформации [3].

Анализ компрометации через обратный инжиниринг предполагает системный подход, включающий изучение архитектурных особенностей моделей, методов их обучения и особенностей данных. Особое внимание уделяется

моделям с высокой емкостью и сложной структурой, таким как трансформеры и глубокие сверточные сети, которые способны «запоминать» специфические детали обучающей выборки, увеличивая риск компрометации. Анализируются уязвимости, возникающие из-за недостаточной регуляризации, отсутствия механизма приватности или ошибок в реализации протоколов обучения и развертывания [4].

Для оценки риска компрометации применяются как формальные методы, так и экспериментальные сценарии. Формальные методы включают построение моделей угроз и определение границ приватности, например, через дифференциальную приватность или теории взаимной информации. Экспериментальные подходы основаны на реализации различных атак на реальные или синтетические модели, что позволяет получить количественные показатели вероятности и масштаба утечки. Важным аспектом является разработка метрик, отражающих не только факт компрометации, но и степень ущерба, потенциально наносимого раскрытием той или иной информации.

Одним из перспективных направлений является интеграция механизмов обнаружения и предотвращения обратного инжиниринга непосредственно в архитектуру и жизненный цикл модели. Это включает применение адаптивных шумовых механизмов, методов агрегации и фильтрации ответов, ограничение доступа к критическим компонентам, а также внедрение процессов регулярного аудита и мониторинга поведения модели. Кроме того, важна разработка протоколов безопасного обмена моделями и обновлениями, особенно в распределённых и федеративных средах, где контроль над данными и их защитой усложнен [5].

Этические и правовые аспекты анализа компрометации конфиденциальных данных через обратный инжиниринг моделей также требуют особого внимания. В условиях растущих требований к защите персональных данных, таких как GDPR и аналогичные международные нормы, обеспечение приватности становится не только технической, но и юридической необходимостью. Создание прозрачных, проверяемых и сертифицируемых процессов анализа и защиты моделей способствует повышению доверия пользователей и регулирующих органов к ИИ-технологиям.

В итоге, анализ компрометации конфиденциальных данных посредством обратного инжиниринга модели представляет собой сложную междисциплинарную задачу, сочетающую технические, организационные и правовые подходы. Глубокое понимание механизмов утечки и разработка эффективных методов защиты – ключ к построению надежных, безопасных и этически ответственных систем искусственного интеллекта, способных эффективно функционировать в условиях современного информационного общества [6].

Список литературы:

1. Chatzikokolakis, K. Broadening the Scope of Differential Privacy Using Metrics / K. Chatzikokolakis, M. E. Andrés, N. E. Bordenabe, C. Palamidessi // Privacy Enhancing Technologies. – Berlin : Springer, 2013. – P. 82–102. – DOI 10.1007/978-3-642-39077-7_5.

2. Duchi, J. C. Local Privacy and Statistical Minimax Rates / J. C. Duchi, M. I. Jordan, M. J. Wainwright // IEEE Annual Symposium on Foundations of Computer Science. – Piscataway : IEEE, 2013. – P. 429–438. – DOI 10.1109/FOCS.2013.53.

3. Wasserman, L. A Statistical Framework for Differential Privacy / L. Wasserman, S. Zhou // Journal of the American Statistical Association. – 2010. – Vol. 105, № 489. – P. 375–389. – DOI 10.1198/jasa.2009.tm08651.

4. Barthe, G. Information-Theoretic Foundations of Differential Privacy / G. Barthe, B. Köpf // IEEE Computer Security Foundations Symposium. – Piscataway : IEEE, 2011. – P. 191–204. – DOI 10.1109/CSF.2011.20.

5. Mardziel, P. Quantifying Information Flow for Dynamic Secrets / P. Mardziel, S. Magill, M. Hicks, M. Srivatsa // IEEE Symposium on Security and Privacy. – Piscataway : IEEE, 2014. – P. 540–555. – DOI 10.1109/SP.2014.41.

6. Boreale, M. Quantitative Information Flow under Generic Leakage Functions and Adaptive Adversaries / M. Boreale, F. Pampaloni, M. Paolini // Journal of Computer Security. – 2015. – Vol. 23, № 2. – P. 167–191. – DOI 10.3233/JCS-140515.