УДК 004.89

ОЦЕНКА КОНФИДЕНЦИАЛЬНОСТИ МОДЕЛЕЙ С ИСПОЛЬЗОВАНИЕМ АТАК НА ИЗВЛЕЧЕНИЕ ДАННЫХ

Эспосито Ф., студент направления Industrial Production Engineering Bachelor's degree programComputer Visions Master's Degree Programm, II курс
Миланский технический университет, г. Милан, Италия

Оценка конфиденциальности моделей искусственного интеллекта с использованием атак на извлечение данных представляет собой важнейшее направление в исследовании безопасности систем машинного обучения. С активным развитием и повсеместным внедрением предобученных моделей в чувствительные сферы – здравоохранение, финансы, государственное администрирование – возросла не только эффективность интеллектуальных решений, но и потенциальные риски, связанные с несанкционированным извлечением информации, использованной в процессе обучения моделей. В отличие от классических векторов угроз, атакующие здесь используют саму модель как источник уязвимости, извлекая из неё приватные данные или статистически значимую информацию об элементах обучающей выборки [1].

Атаки на извлечение данных (data extraction attacks) представляют собой класс методов, направленных на восстановление или генерацию информации, использованной для обучения модели, путём анализа её откликов, внутренних состояний или градиентов. В простейшем варианте, если атакующий имеет доступ к модели и может произвольно задавать ей входные данные (так называемый white-box или gray-box доступ), то он может использовать поведение модели для пошагового восстановления элементов исходной выборки. Например, нейросети с большим числом параметров, склонные к переобучению, особенно подвержены подобным атакам: при недостаточной регуляризации они могут сохранить чрезмерно точные представления об отдельных примерах, что делает возможным их воспроизведение с высокой степенью точности.

Оценка конфиденциальности в таком контексте означает количественную и качественную проверку того, насколько обученная модель уязвима к атакам на извлечение. Одним из важнейших аспектов такой оценки является построение экспериментальных сценариев, моделирующих потенциальные атаки. Применяются методы, основанные на градиентном анализе, контрпримерном поиске, генеративных подходах, а также специальное тестирование модели с использованием метаобучения. Например, если атакующий способен с помощью оптимизационного механизма синтезировать входной пример,

который приводит модель к предсказанию с максимальной уверенностью, и этот пример структурно похож на элемент обучающей выборки, это уже служит признаком потенциальной утечки. Подобные техники используются в атаке Inversion Attack, когда результатом становится частичное или полное восстановление изображения, текста или вектора признаков, из которых состояли обучающие данные [2].

Особую угрозу представляют случаи, когда модель используется в облачном API-доступе. Даже при ограниченном доступе к архитектуре модели (black-box-сценарий), атакующие могут обучить вспомогательную модель-имитатор, собрав множество входов и соответствующих им выходов модели-жертвы. Такой подход, известный как model stealing, может сопровождаться извлечением признаков или даже более глубоким восстановлением индивидуальных наблюдений из обучающего набора, особенно если API выдаёт вероятностные распределения, а не только метки классов. Таким образом, даже высокоуровневое взаимодействие с моделью может быть использовано для вторичной реконструкции чувствительных данных.

Для оценки конфиденциальности модели с учётом таких угроз применяются метрики риска извлечения, такие как extraction precision, extraction recall, membership inference risk и другие показатели, отражающие вероятность того, что конкретное наблюдение из выборки было «узнано» моделью и поддаётся реконструкции. Один из наиболее широко используемых инструментов – тесты на включённость в выборку (membership inference attacks), в которых проверяется, может ли атакующий отличить, использовался ли конкретный пример в обучении модели. Если такая классификация работает лучше случайного угадывания, модель считается уязвимой. Подобные тесты показывают, что даже большие языковые модели, такие как GPT или BERT, могут сохранять в своём параметрическом пространстве признаки конкретных строк, документов или записей, если не были предприняты специальные меры по защите приватности [3].

Важным направлением защиты от атак на извлечение данных является внедрение механизмов дифференциальной приватности (DP). При использовании DP в процессе обучения к градиентам или выходным данным добавляется математически обоснованный шум, который снижает вероятность того, что модель сохранит конкретные детали об отдельных наблюдениях. Однако применение DP требует тонкой настройки: при слишком сильном шуме модель теряет обобщающую способность, а при слишком слабом – защита становится неэффективной. Кроме DP, также применяются техники регуляризации (например, dropout, weight decay), обрезки градиентов и замены точных предсказаний на дискретизированные, что снижает информативность отклика модели и затрудняет обратный анализ [4].

Наконец, оценка конфиденциальности модели должна быть интегрирована в цикл её жизненного управления. Это включает в себя автоматизированное тестирование модели на уязвимости, регулярный аудит по стандартам безопасности ИИ, отслеживание статистики доступа, а также использование генеративных тестов, направленных на имитацию атак. Особенно важно проводить такие проверки при публикации предобученных моделей в открытом доступе, так как пользователи могут применять их в конфиденциальных сценариях, не осознавая степень риска. При отсутствии механизмов оценки и ограничения извлекаемости данных такие модели становятся не только вычислительным инструментом, но и потенциальным источником утечки данных.

Таким образом, оценка конфиденциальности моделей ИИ с использованием атак на извлечение данных является междисциплинарной задачей, объединяющей элементы теории информации, оптимизации, безопасности и прикладного анализа ИИ. Эффективное применение таких методов позволяет не только выявлять риски и слабые места в архитектуре моделей, но и закладывать прочный фундамент для создания доверенного, приватного и этичного искусственного интеллекта [5].

Список литературы:

1. Hayes, J. LOGAN: Membership Inference Attacks Against Generative Models / J. Hayes, L. Melis, G. Danezis, E. De Cristofaro // Proceedings on Privacy Enhancing Technologies. – 2019. – № 1. – P. 133–152. – DOI 10.2478/popets-2019-0008.

2. Chen, D. GAN-Leaks: A Taxonomy of Membership Inference Attacks against Generative Models / D. Chen, N. Yu, Y. Zhang, M. Fritz // Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. – New York : ACM, 2020. – P. 343–362. – DOI 10.1145/3372297.3417238.

3. Hitaj, B. Deep Models Under the GAN: Information Leakage from Collaborative Deep Learning / B. Hitaj, G. Ateniese, F. Perez-Cruz // Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. – New York : ACM, 2017. – P. 603–618. – DOI 10.1145/3133956.3134012.

4. Melis, L. Exploiting Unintended Feature Leakage in Collaborative Learning / L. Melis, C. Song, E. De Cristofaro, V. Shmatikov // IEEE Symposium on Security and Privacy. – Piscataway : IEEE, 2019. – P. 691–706. – DOI 10.1109/SP.2019.00029.

5. Zhu, L. Deep Leakage from Gradients / L. Zhu, Z. Liu, S. Han // Advances in Neural Information Processing Systems. – 2019. – Vol. 32. – P. 14774–14784.