УДК 004.89

АНАЛИЗ УЯЗВИМОСТЕЙ МОДЕЛЕЙ ИИ К АТАКАМ С ВОССТАНОВЛЕНИЕМ ОБУЧАЮЩЕЙ ВЫБОРКИ

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

Анализ уязвимостей моделей искусственного интеллекта (ИИ) к атакам с восстановлением обучающей выборки представляет собой важную и сложную задачу в области безопасности машинного обучения. В последнее десятилетие использование ИИ в критически важных сферах, таких как здравоохранение, финансы и безопасность, стало нормой, однако с этим возникли новые угрозы, связанные с возможностью утечки конфиденциальной информации через такие модели [1].

Одна из таких угроз – атаки с восстановлением обучающей выборки, при которых злоумышленники пытаются извлечь данные, использованные для обучения модели, что может привести к утечке личных данных, нарушению конфиденциальности или даже ухудшению функционирования системы.

Для понимания масштабов этой угрозы необходимо исследовать как происходят такие атаки, какие уязвимости в моделях ИИ способствуют их успешному проведению и какие меры можно предпринять для защиты.

Атаки с восстановлением обучающей выборки основываются на идее, что даже после завершения обучения, модель может хранить в своих параметрах информацию, которая позволяет восстановить или реконструировать часть данных, на которых она была обучена. Это особенно опасно, если модель обучалась на чувствительных или личных данных, например, медицинских записях, финансовых транзакциях или персональных данных клиентов. Модели ИИ, такие как нейронные сети, могут unintentionally «запоминать» информацию о специфических примерах из обучающей выборки в своих весах и активациях, что может быть извлечено при соответствующих атаках [2].

Одним из наиболее известных видов атак на ИИ-модели является атакующий доступ к внутренним состояниям модели – то есть попытка извлечь информацию, скрытую в весах или активациях модели, которая может быть использована для восстановления обучающих данных. Это может происходить через использование градиентных методов, таких как обратная пропаганда (backpropagation), при которых атакующий анализирует поведение модели на

новых примерах, чтобы определить, как она реагирует на изменения входных данных.

Такой анализ позволяет построить гипотезы о содержимом обучающей выборки, и, на основе этих гипотез, можно восстановить данные или их особенности.

Для эффективного анализа уязвимостей модели к атакам с восстановлением обучающей выборки необходимо использовать методы, позволяющие оценить чувствительность модели к атакующим воздействиям.

Оценка чувствительности в данном контексте означает проверку, насколько изменение входных данных влияет на результат предсказания модели. Если модель демонстрирует большую зависимость от отдельных примеров из обучающего набора, это может указывать на высокую вероятность того, что такие примеры могут быть восстановлены или экстраполированы злоумышленниками. Это особенно важно для моделей, использующих сложные архитектуры, такие как глубокие нейронные сети, где скрытые слои могут скрывать достаточно много информации о входных данных, что создает риски утечек [3].

Для диагностики уязвимости моделей к восстановлению обучающих данных применяется несколько методов. Один из них – это атакующие методы на основе обратной пропаганды градиентов, которые используются для анализа, какие примеры из обучающей выборки могли оказать наибольшее влияние на веса и активации модели. Эти методы позволяют атакующему вычислить, какие конкретные данные влияли на решение модели в наибольшей степени, и восстановить их на основе градиентов, вычисленных для текущего ввода.

Другим подходом является генеративный анализ с использованием генеративных моделей – например, Generative Adversarial Networks (GAN). В этих атаках модель пытается генерировать примеры данных, которые максимально похожи на те, что использовались для её обучения.

Генеративные модели, обученные на результатах работы целевой модели, могут быть использованы для извлечения информации о характеристиках обучающих данных и реконструкции их структуры. Это может привести к восстановлению личных данных или других чувствительных характеристик, что увеличивает риск утечек [4].

Кроме того, с целью уменьшения рисков утечек разработаны методы дифференциальной приватности, которые помогают ограничить степень запоминания модели обучающих данных. Применяя технику дифференциальной приватности, можно добавить шум к данным или результатам обучения таким образом, чтобы извлечение точной информации из модели становилось невозможным. Дифференциальная приватность использует параметры, такие как

уровень шума (ε), для контролирования того, насколько модель «запоминает» или раскрывает информацию о данных.

Например, применение дифференциально приватных алгоритмов при обучении модели может значительно затруднить восстановление обучающих данных, поскольку шум делает их трудными для точной реконструкции.

Важным аспектом защиты от атак с восстановлением обучающих выборок является анализ метаданных модели, таких как её архитектура, использование регуляризации, а также параметры оптимизации. Модели, которые используют низкий уровень регуляризации или имеют большое количество параметров, могут быть более уязвимыми к восстановлению данных. Применение методов регуляризации, таких как L2-регуляризация или dropout, может уменьшить вероятность того, что модель будет сильно зависеть от отдельных примеров в обучающем наборе и, следовательно, снизить риски утечек.

Для улучшения безопасности моделей ИИ также важно проводить периодическое тестирование и аудит на уязвимости. Это включает в себя применение специализированных инструментов, которые могут симулировать атаки с восстановлением данных на основе различных методик и проверять, насколько защищена модель. Параллельно с этим, важно внедрять системы мониторинга и отслеживания поведения модели в реальных условиях, что поможет выявить аномалии, связанные с возможными утечками данных [5].

Для заключения, атаки с восстановлением обучающей выборки являются серьезной угрозой для безопасности моделей искусственного интеллекта, поскольку могут привести к утечке конфиденциальной информации.

Методы диагностики уязвимостей в таких моделях включают анализ чувствительности, использование генеративных моделей для симуляции атак, а также внедрение дифференциальной приватности и регуляризации.

Важно проводить регулярные тесты и аудиты на уязвимости, чтобы убедиться, что модель защищена от подобных атак. С развитием технологий и ростом использования ИИ в различных областях, обеспечение безопасности моделей и защита от утечек данных становятся важнейшими приоритетами для разработчиков и исследователей в этой области [6].

Список литературы:

1. Yeom, S. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting / S. Yeom, I. Giacomelli, M. Fredrikson, S. Jha // IEEE Computer Security Foundations Symposium. – Piscataway : IEEE, 2018. – P. 268–282. – DOI 10.1109/CSF.2018.00027.

2. Salem, A. ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models / A. Salem, Y. Zhang, M.

Humbert [et al.] // Network and Distributed System Security Symposium. – San Diego : Internet Society, 2019. – P. 1–15.

3. Long, Y. Understanding Membership Inferences on Well-Generalized Learning Models / Y. Long, V. Bindschaedler, L. Wang [et al.] // arXiv preprint. – 2018. – arXiv:1802.04889.

4. Truex, S. Demystifying Membership Inference Attacks in Machine Learning as a Service / S. Truex, L. Liu, M. E. Gursoy [et al.] // IEEE Transactions on Services Computing. – 2021. – Vol. 14, № 6. – P. 2073–2089. – DOI 10.1109/TSC.2019.2897554.

5. Sablayrolles, A. White-Box vs Black-Box: Bayes Optimal Strategies for Membership Inference / A. Sablayrolles, M. Douze, C. Schmid [et al.] // Proceedings of the 36th International Conference on Machine Learning. – 2019. – Vol. 97. – P. 5558–5567.

6. Hilprecht, B. Monte Carlo and Reconstruction Membership Inference Attacks against Generative Models / B. Hilprecht, M. Härterich, D. Bernau // Proceedings on Privacy Enhancing Technologies. – 2019. – № 4. – P. 232–249. – DOI 10.2478/popets-2019-0067