УДК 004.89

МЕТОДЫ ОЦЕНКИ РИСКА УТЕЧКИ ДАННЫХ В ПРЕДОБУЧЕННЫХ МОДЕЛЯХ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Кобелев М.В., студент гр. ИТб-221, III курс
Научный руководитель: Николаева Е.А., к. ф.-м. н., доцент
Кузбасский государственный технический
университет имени Т.Ф. Горбачева, г. Кемерово

Методы оценки риска утечки данных в предобученных моделях искусственного интеллекта (ИИ) становятся важным элементом обеспечения безопасности и конфиденциальности в современных системах машинного обучения. С увеличением применения ИИ в различных сферах, таких как здравоохранение, финансы, юриспруденция и другие критически важные области, возрастает и угроза утечек данных, что может иметь далеко идущие последствия для пользователей и организаций [1].

Важно отметить, что даже предобученные модели, которые изначально могут не использовать чувствительные данные напрямую, все равно могут случайно «выучить» и передавать скрытые закономерности, которые позволяют извлекать конфиденциальную информацию о частных лицах. Оценка и предотвращение таких утечек являются неотъемлемой частью разработки безопасных и этичных моделей ИИ.

Оценка риска утечки данных в предобученных моделях ИИ начинается с анализа источников утечек. Одним из таких источников является избыточная информация, хранящаяся в модели, которая может быть извлечена с помощью различных техник, таких как обратная инженерия или методы извлечения информации, использующие доступ к параметрам модели.

Например, в случае нейронных сетей, обученных на медицинских данных, скрытые слои модели могут случайно запомнить и сохранить информацию, которая позволяет воссоздать части исходных данных. Это может стать угрозой, если такие модели используются в открытых или частных приложениях, и внешние лица могут получить доступ к этим данным через атаку или обратный анализ.

Для оценки риска утечки данных применяются методы анализа чувствительности модели. Суть этого подхода заключается в том, чтобы проверить, насколько сильно изменения в входных данных могут повлиять на выходные предсказания модели. Если модель демонстрирует большую чувствительность к отдельным признакам или небольшим изменениям в данных, это может

означать, что она хранила слишком много информации о чувствительных атрибутах, что в свою очередь увеличивает риск утечки [2].

Методы обратной пропаганды и градиентного спуска могут использоваться для оценки, какие входные данные (например, личная информация) могут привести к утечке через изменение параметров модели.

Одним из ключевых подходов в оценке риска утечек является дифференциальная приватность. Это метод защиты данных, который добавляет шум к данным или результатам модели таким образом, чтобы не было возможным извлечь информацию о любом конкретном примере в обучающих данных.

Применение дифференциальной приватности в предобученных моделях помогает минимизировать риски утечек данных, даже если модель подвергается атаке, направленной на извлечение чувствительной информации. Оценка дифференциальной приватности модели может быть выполнена с помощью анализа метрик приватности, таких как ε-дифференциальная приватность, где меньшие значения ε указывают на более высокий уровень защиты [3].

Другим методом является тестирование модели на утечку с использованием экспериментальных атак, например, атак с восстановлением данных. В таких атаках атакующие пытаются извлечь информацию из модели, выдавая на вход модель известные данные и наблюдая, как она реагирует.

Если модель выдает результаты, которые могут быть использованы для восстановления или реконструкции исходных данных, это сигнализирует о наличии утечки. Это могут быть как атаки на традиционные нейронные сети, так и более сложные системы, такие как трансформеры или модели с обучением на больших данных. Такие методы позволяют напрямую оценить, насколько предобученная модель уязвима к утечкам и какие данные могут быть извлечены.

Анализ избыточных зависимостей в предобученных моделях также является важным шагом в оценке риска утечек. Модели, которые слишком сильно зависят от определённых признаков, могут перенести скрытую информацию, которая может быть использована для восстановления конфиденциальных данных. Для анализа таких зависимостей могут быть использованы методы снижения размерности, такие как PCA (Principal Component Analysis) или t-SNE (t-Distributed Stochastic Neighbor Embedding), чтобы выявить, какие компоненты или признаки из обучающих данных имеют наибольшее влияние на модель и могут привести к утечкам. Эти методы помогают уменьшить размерность данных и выявить скрытые зависимости, которые могут быть использованы для предсказания чувствительных атрибутов [4].

Кроме того, на этапе оценки риска утечки данных важно учесть возможность переноса знаний между моделями. В случае, если модель была обучена на одной задаче и затем перенесена на другую (например, для решения новой

задачи на другом наборе данных), важно провести оценку, не перенёсла ли модель информацию, которая может быть конфиденциальной в контексте исходной задачи. Методики, такие как fine-tuning (доработка) на новых данных, могут привести к утечке информации, если изначальная модель имела доступ к чувствительным данным. Важно тестировать переносимость модели и оценивать риск утечек при изменении контекста её применения.

Также существует подход с использованием генеративных моделей для диагностики утечек. Генеративные модели, такие как GAN (Generative Adversarial Networks) или VAEs (Variational Autoencoders), могут быть использованы для создания синтетических данных на основе предобученной модели, и анализировать, насколько эти синтетические данные похожи на исходные, конфиденциальные данные. Если модель генерирует данные, которые слишком близки к реальным данным, это указывает на возможность утечек и повышает риск.

В заключение, методы оценки риска утечки данных в предобученных моделях искусственного интеллекта играют ключевую роль в защите конфиденциальности и обеспечении безопасности. Эти методы помогают выявить и минимизировать риски утечек данных, обеспечивая доверие к моделям и их результатам.

С использованием таких подходов, как анализ чувствительности, дифференциальная приватность, тестирование на утечку, анализ избыточных зависимостей и генеративные модели, можно значительно повысить уровень безопасности предобученных моделей и сделать их более защищёнными от возможных угроз утечек данных [5].

Список литературы:

1. Carlini, N. Extracting Training Data from Large Language Models / N. Carlini, F. Tramer, E. Wallace [et al.] // Proceedings of the 30th USENIX Security Symposium. – Berkeley : USENIX Association, 2021. – P. 2633–2650.

2. Carlini, N. The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks / N. Carlini, C. Liu, U. Erlingsson [et al.] // Proceedings of the 28th USENIX Security Symposium. – Berkeley : USENIX Association, 2019. – P. 267–284.

3. Carlini, N. Membership Inference Attacks From First Principles / N. Carlini, S. Chien, M. Nasr [et al.] // IEEE Symposium on Security and Privacy. – Piscataway : IEEE, 2022. – P. 1897–1914. – DOI 10.1109/SP46214.2022.9833649.

4. Carlini, N. Extracting Training Data from Diffusion Models / N. Carlini, J. Hayes, M. Nasr [et al.] // Proceedings of the 32nd USENIX Security Symposium. – Berkeley : USENIX Association, 2023. – P. 5253–5270.

5. Nasr, M. Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-Box Inference Attacks against Centralized and Federated Learning / M. Nasr, R. Shokri, A. Houmansadr // IEEE Symposium on Security and Privacy. – Piscataway : IEEE, 2019. – P. 739–753. – DOI 10.1109/SP.2019.00065.