УДК 004.89

ИНСТРУМЕНТЫ АУДИТА ПРЕДОБУЧЕННЫХ МОДЕЛЕЙ В УСЛОВИЯХ ОГРАНИЧЕННОГО ДОСТУПА К ДАННЫМ

Чиков А.И., Оператор научной роты, III курс
Военная академия связи имени Маршала Советского Союза С. М. Буденного,
г. Санкт-Петербург

По мере интеграции предобученных моделей машинного обучения в прикладные задачи растёт потребность в их аудите – систематической проверке корректности, надёжности, устойчивости и соответствия нормативным требованиям. Однако одной из основных проблем, с которой сталкиваются разработчики и исследователи в реальной практике, является ограниченность доступа к данным, как по техническим, так и по юридическим причинам.

Это может быть связано с конфиденциальностью персональной информации (например, в здравоохранении и юриспруденции), с коммерческой тайной (например, в индустриальных системах и финансах), либо с условиями лицензирования самой модели. В такой ситуации классические методы тестирования и интерпретации оказываются неприменимыми, что требует разработки и внедрения специализированных инструментов аудита предобученных моделей в условиях ограниченного доступа к данным [1].

Цель аудита в данном контексте заключается не просто в том, чтобы оценить точность модели или её архитектуру, а в том, чтобы проверить, как модель ведёт себя в неизвестных или потенциально опасных условиях, не имея полного доступа к обучающим или операционным данным. В идеале, такой аудит должен быть независимым, воспроизводимым и соответствующим требованиям прозрачности, особенно если модель используется в социально чувствительных или регулируемых сферах [2].

Одним из ключевых направлений разработки является создание инструментов «чёрного ящика» (black-box auditing), позволяющих взаимодействовать с моделью исключительно через её входы и выходы. Эти инструменты не требуют знания архитектуры или доступа к весам, что делает их применимыми даже к проприетарным API, таким как коммерческие языковые модели или классификаторы [3].

Среди типовых подходов:

- Методы тестирования на обобщённость: используется синтетически созданный набор входов, покрывающий репрезентативное множество сценариев. Оценивается стабильность и

консистентность выходов. Такие подходы, как CheckList, позволяют строить матрицы функционального покрытия и выявлять "дырки" в логике модели [4].

- Оценка чувствительности к пертурбациям: даже без доступа к данным, можно создавать небольшие модификации входов (например, через перестановку слов, изменение структуры, добавление шума) и оценивать, насколько результат модели устойчив к таким изменениям. Это даёт косвенную информацию о надёжности и робастности модели [5].
- Анализ согласованности: сравниваются ответы модели на семантически эквивалентные, но формально различающиеся запросы. При высокой чувствительности к формулировке можно заподозрить либо переобучение, либо недостаточную обоснованность предсказаний [6].

Другой важный класс инструментов опирается на методы метаанализа модели без доступа к данным, но с доступом к эмбеддингам или к логам предсказаний. Например, можно использовать:

- Метрики разнообразия выходов: в условиях многократного запуска модели с вариациями в входах можно статистически оценить разброс выходов, что сигнализирует о нестабильности.
- Оценка вероятностной уверенности: модели, особенно предобученные трансформеры, возвращают вероятностные оценки. Аномально высокие или низкие вероятности на очевидных примерах могут указывать на проблемы с калибровкой модели.

В условиях ограниченного доступа также развивается направление zero-knowledge auditing – аудит без доступа к данным и без раскрытия внутренней логики модели, при этом с сохранением возможности верификации. Такие подходы опираются на криптографические протоколы (например, ZK-SNARKs) или на доверенные вычисления (trusted execution environments, TEE), что особенно актуально при аудите модели сторонней организацией.

При использовании моделей в чувствительных отраслях всё чаще применяется инструментальный аудит на основе сгенерированных тестов, при этом генерация может быть реализована на основе малых открытых датасетов, эталонных сценариев или даже с помощью другой модели.

Такой подход позволяет обойти ограничения доступа к реальным пользовательским данным, не нарушая конфиденциальности. Так, например, в медицине могут использоваться синтетические кейсы пациентов, в финансах – псевдооперации, в промышленности – симуляции производственных процессов.

Особую нишу занимают аудит-инструменты, встраиваемые в продуктивные пайплайны, которые собирают анонимизированные и агрегированные

метаданные о работе модели. Это могут быть частотные характеристики входов, статистика распределений выходов, частота отклонений и ошибки на валидационных подсэмплах.

Несмотря на то, что прямой доступ к данным при этом отсутствует, анализ такого рода телеметрии позволяет с высокой степенью достоверности судить о возможных проблемах и отклонениях в поведении модели. Такие подходы особенно востребованы в банковской сфере, страховании, медицине и облачных системах.

При проектировании инструментов аудита в условиях ограниченного доступа важна юридическая и нормативная совместимость. Аудит должен не только выявлять риски и дефекты, но и обеспечивать возможности для отчётности перед регуляторами и заказчиками, включая возможность документировать процесс и сохранять следы верификации.

Это требует интеграции инструментов аудита с системами версионирования, управления метаданными и CI/CD инфраструктурой.

В перспективе именно такие инструменты станут основой доверенного ИИ, способного соответствовать принципам ответственности, проверяемости и непредвзятости. Становится очевидным, что в условиях ограниченного доступа к данным нельзя полагаться исключительно на точечные метрики качества.

Требуется создание экосистемы инструментов аудита, работающих на разных уровнях – от поведения модели в интерактивных сессиях до агрегированной статистики использования и отклонений.

В заключение, инструменты аудита предобученных моделей в условиях ограниченного доступа к данным – это ключевой компонент зрелой и этически обоснованной системы машинного обучения.

Их развитие позволяет не только выявлять потенциальные угрозы и ошибки, но и восстанавливать прозрачность и подотчётность в условиях «чёрного ящика», что критически важно как для разработчиков, так и для пользователей, надзорных органов и общества в целом.

Список литературы:

1. Tramèr, F. Stealing Machine Learning Models via Prediction APIs / F. Tramèr, F. Zhang, A. Juels [et al.] // Proceedings of the 25th USENIX Security Symposium. – Berkeley : USENIX Association, 2016. – P. 601–618.

2. Orekondy, T. Knockoff Nets : Stealing Functionality of Black-Box Models / T. Orekondy, B. Schiele, M. Fritz // 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. – Long Beach : IEEE, 2019. – P. 4949–4958. – DOI 10.1109/CVPR.2019.00509.

3. Juuti, M. PRADA : Protecting against DNN Model Stealing Attacks / M. Juuti, S. Szyller, S. Marchal, N. Asokan // 2019 IEEE European Symposium on Security and Privacy. – Stockholm : IEEE, 2019. – P. 512–527. – DOI 10.1109/EuroSP.2019.00044.

4. Jagielski, M. High Accuracy and High Fidelity Extraction of Neural Networks / M. Jagielski, N. Carlini, D. Berthelot [et al.] // Proceedings of the 29th USENIX Security Symposium. – Berkeley : USENIX Association, 2020. – P. 1345–1362.

5. 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.

6. 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 // 2019 IEEE Symposium on Security and Privacy. – San Francisco : IEEE, 2019. – P. 739–753. – DOI 10.1109/SP.2019.00065.