УДК 004.89

АНАЛИЗ УЯЗВИМОСТЕЙ ПРЕДОБУЧЕННЫХ МОДЕЛЕЙ И ИНСТРУМЕНТЫ ДЛЯ ИХ НЕЙТРАЛИЗАЦИИ

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

Предобученные модели, особенно крупные трансформерные архитектуры и глубокие нейросетевые системы, сегодня становятся основой большинства современных приложений искусственного интеллекта. Однако вместе с их ростом по размеру и универсальности увеличивается и поверхность атаки – количество потенциальных уязвимостей, которые могут быть использованы злоумышленниками, случайными или целенаправленными действиями, приводящими к некорректному или опасному поведению модели.

Особенно это актуально в условиях, когда такие модели применяются в автоматизированных решениях в медицине, праве, финансовой аналитике, образовании и цифровом управлении. Поэтому необходимым направлением в построении доверенного ИИ является анализ уязвимостей предобученных моделей и создание инструментов, способных их нейтрализовать или компенсировать на разных этапах жизненного цикла модели – от проектирования и валидации до эксплуатации и деактивации.

Прежде всего, важно различать несколько типов уязвимостей. Один из наиболее изученных классов – адверсариальные атаки, при которых минимальные (и зачастую невидимые для человека) изменения входа приводят к существенным искажениям вывода. В языковых моделях это может быть подмена синонима, добавление пунктуации, изменения регистра, перестановка слов, в результате чего модель теряет семантическую устойчивость и генерирует ошибочный или токсичный вывод. В компьютерном зрении аналогичные искажения могут представлять собой небольшие шумы, градиентные паттерны или изменение освещения. Такие атаки показывают, насколько нелинейной и хрупкой может быть поверхность принятия решений у даже очень мощной модели [1].

Второй важный класс – атаки на обучение (training-time attacks), например, data poisoning, при котором злоумышленник внедряет в тренировочные данные специально сконструированные примеры, способные внедрить «логические закладки» в поведение модели. Это может проявляться в виде скрытых правил активации – например, модель, которая классифицирует текст как

позитивный, если в нём встречается определённая бессмысленная последовательность символов. Такая атака может быть использована для обхода фильтров, цензуры или защиты в системах модерации и категоризации [2].

Отдельного внимания заслуживают атаки на приватность, такие как модельные инверсии и атакующие генерации, позволяющие по выходу модели восстановить исходные входные данные или даже извлечь части обучающего корпуса. Это особенно опасно в системах, работающих с чувствительной информацией (например, медицинские отчёты или переписка). Механизмы генерации в крупных языковых моделях могут непреднамеренно «вспоминать» и повторять фразы, имена или события из приватного обучающего контекста, если защита данных не была реализована на этапе подготовки корпуса [3].

В этом контексте инструменты анализа уязвимостей предобученных моделей начинают играть двойную роль: с одной стороны, они выявляют и классифицируют слабые места, а с другой – становятся основой для проектирования механизмов нейтрализации и устойчивости.

Один из центральных подходов – автоматизированное тестирование на устойчивость, когда модели подвергаются систематическим и воспроизводимым атакам с помощью библиотек, таких как TextAttack, AdvGLUE, RobustBench, CleverHans, Foolbox. Эти библиотеки предоставляют как стандартные атаки, так и возможность задавать кастомные сценарии для конкретной модели. Анализируется чувствительность к определённым входам, способность распознавать вредоносные паттерны, устойчивость к переформулировке или перенормализации входа. Результаты тестов формируют карту уязвимостей, на основе которой выстраиваются стратегии компенсации [4].

Другим важным инструментом являются фреймворки для интерпретации логики модели, такие как Captum, SHAP, LIME, которые позволяют визуализировать внутренние зависимости между входами и выходом, выявлять неожиданные корреляции, «спусковые крючки» и переобученные шаблоны. Особенно это актуально для выявления закодированных зависимостей от неочевидных признаков, таких как стилистика, акценты, вторичные фоновые объекты. Если модель, к примеру, классифицирует опухоль по яркости фона снимка, это может быть выявлено через heatmap или анализ градиентов. Таким образом, интерпретация становится не только средством объяснения, но и способом идентификации опасных локальных решений.

Для нейтрализации уязвимостей используется несколько категорий решений:

1. Защитное дообучение (adversarial training), при котором модель обучается на «усиленном» наборе данных, включающем как нормальные, так и атакующие примеры. Это позволяет модели формировать более устойчивые границы принятия решений.

2. Превентивная фильтрация входов, когда до передачи данных в модель осуществляется синтаксический, семантический или визуальный анализ на предмет подозрительных искажений. Такой подход эффективен, если атака осуществляется извне.
3. Ограничение доступа к чувствительным внутренним слоям модели, особенно в облачных решениях, где API может быть использован для зондирования модели. Это снижает риск обратной инженерии и инверсии [5].
4. Дифференциальная приватность, обеспечивающая, чтобы вклад отдельного примера в обучение модели не мог быть восстановлен даже при полной доступности параметров. Библиотеки, такие как Opacus и Google DP, позволяют реализовать такие меры при fine-tuning модели [6].
5. Аудит поведения модели в реальном времени, при котором отслеживаются отклонения от типичного распределения выходов, аномальные паттерны запросов и повторяющиеся ошибки. Это позволяет выявлять как пользовательские атаки, так и внутренние нарушения [7].
6. Регуляризация доверительных границ, при которой модель при высоком уровне неопределенности вынуждена отказываться от ответа или запрашивать подтверждение от человека-оператора.

Таким образом, анализ уязвимостей и инструменты их нейтрализации составляют неотъемлемую часть архитектуры доверенного ИИ, особенно в эпоху массового внедрения моделей в социально чувствительные и регламентируемые области. В условиях, когда каждая ошибка модели может нести этические, юридические или финансовые последствия, способность заранее обнаруживать и устранять потенциальные векторы атаки становится основой зрелой и ответственной инженерной практики.

Необходимо формировать культуру предсказуемости и прозрачности в дизайне моделей: каждая предобученная система должна сопровождаться картой рисков, журналом известных уязвимостей и планом реагирования на нарушения. Лишь в этом случае искусственный интеллект сможет интегрироваться в критическую инфраструктуру без потери доверия со стороны пользователей, организаций и регуляторов.

Список литературы:

1. Goodfellow, I. J. Explaining and Harnessing Adversarial Examples / I. J. Goodfellow, J. Shlens, C. Szegedy // International Conference on Learning Representations. – 2015.

2. Szegedy, C. Intriguing Properties of Neural Networks / C. Szegedy, W. Zaremba, I. Sutskever [et al.] // International Conference on Learning Representations. – 2014.

3. Biggio, B. Evasion Attacks against Machine Learning at Test Time / B. Biggio, I. Corona, D. Maiorca [et al.] // Machine Learning and Knowledge Discovery in Databases. – Berlin : Springer, 2013. – P. 387–402. – DOI 10.1007/978-3-642-40994-3_25.

4. Morris, J. X. TextAttack : A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP / J. X. Morris, E. Lifland, J. Y. Yoo [et al.] // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations. – Stroudsburg : ACL, 2020. – P. 119–126. – DOI 10.18653/v1/2020.emnlp-demos.16.

5. Fredrikson, M. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures / M. Fredrikson, S. Jha, T. Ristenpart // Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. – New York : ACM, 2015. – P. 1322–1333. – DOI 10.1145/2810103.2813677.

6. Abadi, M. Deep Learning with Differential Privacy / M. Abadi, A. Chu, I. Goodfellow [et al.] // Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. – New York : ACM, 2016. – P. 308–318. – DOI 10.1145/2976749.2978318.

7. Nicolae, M.-I. Adversarial Robustness Toolbox v1.0.0 / M.-I. Nicolae, M. Sinn, M. N. Tran [et al.] // arXiv. – 2018. – arXiv:1807.01069.