УДК 004.89

МЕТОДЫ ОЦЕНКИ УЯЗВИМОСТЕЙ МОДЕЛЕЙ ИИ К АТАКУЮЩИМ СТРАТЕГИЯМ ПО АНАЛОГИИ

Лукашенок В.И., младший научный сотрудник
Военная академия связи имени Маршала Советского Союза С. М. Буденного,
г. Санкт-Петербург

Оценка уязвимостей моделей искусственного интеллекта к атакующим стратегиям по аналогии представляет собой сравнительно новое, но быстро развивающееся направление, направленное на изучение устойчивости ИИ-систем к атакам, не совпадающим буквально с обучающими данными атакующего, но воспроизводящим их стратегические, структурные или поведенческие особенности. В отличие от классических атак с прямым подбором входов или градиентными методами (например, FGSM, PGD), атаки по аналогии используют сходство в концептуальном или функциональном пространстве между различными классами данных или между разными моделями. Это делает подобные атаки особенно опасными в реальных условиях, где прямой доступ к обучающим данным или к внутренним механизмам модели отсутствует, но поведенческое сходство можно восстановить за счёт анализа контекста и окружающих паттернов [1].

Атаки по аналогии опираются на идею, что многие модели, особенно крупные предобученные нейросети, демонстрируют общие уязвимости в отношении групп семантически или структурно близких входов. Например, если модель уязвима к атакующим примерам, сгенерированным на базе изображений кошек, то она с высокой вероятностью будет аналогично уязвима к атакам, построенным на схожих по структуре объектах – собаках, лисах или даже меховых предметах несмотря на то, что конкретные представления ранее не использовались в обучении. Подобный эффект наблюдается также в языковых моделях: если модель склонна выдавать токсичные или политически окрашенные ответы на определённые темы (например, касающиеся миграции или религии), то атаки по аналогии могут подменять слова синонимами, менять порядок фраз, использовать нейтральные эвфемизмы – и всё равно добиваться нежелательных реакций.

Методы оценки таких уязвимостей включают как эмпирические, так и теоретические подходы. Одним из важнейших направлений является анализ обобщения атак. Здесь оценивается способность атак, сгенерированных на одной выборке или модели, переноситься на другую выборку или архитектуру.

Например, если adversarial-примеры, построенные для одной модели ResNet, успешно нарушают предсказания другой, архитектурно отличной модели ViT, это свидетельствует о наличии обобщающей уязвимости, основанной не на деталях архитектуры, а на принципиальных чертах пространств признаков. В рамках атак по аналогии изучается способность таких «переносных» атак работать в «серых» и «чёрных» коробках (gray-box, black-box), где атакующий не имеет прямого доступа ни к архитектуре, ни к обучающим данным модели-жертвы [2].

Для количественной оценки уязвимостей применяются различные метрики: степень падения точности на тестовой выборке после применения атак, значение метрик перекрытия (например, Jaccard или cosine similarity) между исходными и аналогичными примерами, плотность кластеров атакуемых точек в латентном пространстве модели, а также чувствительность градиентов к малым деформациям семантических признаков. Одним из перспективных методов является использование контрфактического моделирования, когда к одному и тому же примеру применяется множество семантически или функционально эквивалентных трансформаций, и оценивается, насколько изменяются предсказания модели. Большая изменчивость при минимальных трансформациях указывает на слабую устойчивость к стратегиям по аналогии.

Особое внимание в данной области уделяется семантически согласованным атакам, которые сохраняют смысл и читаемость (в текстах) или форму и контекст (в изображениях), но всё равно приводят к ошибке. В компьютерном зрении это могут быть перерисованные контуры объектов, изменения освещения, частичная окклюзия, добавление фонов, маскировка объектов под визуально схожие – иными словами, атаки, апеллирующие к общим когнитивным механизмам восприятия, а не к пиксельным искажениями [3].

В языковом ИИ такие атаки включают переформулировку вопросов, использование парафраз, редактирование грамматики, замену именованных сущностей на аналогичные по роли, но не по значению.

Методология построения и оценки устойчивости к стратегиям по аналогии требует создания специальных тестовых наборов, включающих пары или группы данных с контролируемыми аналогиями. Это может быть достигнуто через синтетические данные, генеративные модели (например, с использованием GPT или diffusion-сетей), а также с помощью алгоритмов семантического сопоставления, таких как WordNet, BERTScore или метод ближайших соседей в эмбеддинг-пространствах. Построенные наборы позволяют эмпирически тестировать модель на способность устойчиво интерпретировать различные формы одного и того же запроса или представления [4].

Важно отметить, что устойчивость к атакам по аналогии критически важна в задачах реального времени и прикладного использования, таких как

автономное вождение, медицинская диагностика, юридическая аналитика, модерация контента, где ошибки на «пограничных» случаях могут иметь фатальные последствия. Модель, успешно прошедшая классическое тестирование на шумы или искажения, может быть полностью дезориентирована, если ей предъявить пример, который она «не ожидала», но который тем не менее полностью эквивалентен тем, что она уже видела – только под иным углом, в иной форме или с иным эмоциональным окрасом.

Наконец, одним из перспективных направлений является разработка мета-моделей оценки уязвимостей, способных автоматически предсказывать слабые места основной модели, обучаясь на истории атак и успешных аналогий. Такие мета-модели могут использоваться для автоматической генерации тестов, уточнения стратегий регуляризации или формирования более устойчивых архитектур. Таким образом, методы оценки уязвимостей к стратегиям по аналогии представляют собой важнейший элемент контуров доверия и безопасности ИИ, обеспечивая не только реактивную диагностику, но и проактивную защиту систем от атак, которые не повторяют известные примеры, но стратегически их воспроизводят [5].

Список литературы:

1. Anil, R. Large-Scale Differentially Private BERT / R. Anil, B. Ghazi, V. Gupta [et al.] // International Conference on Learning Representations. – 2022. – P. 1–24.

2. Kurakin, A. Toward Training at ImageNet Scale with Differential Privacy / A. Kurakin, S. Song, S. Chien [et al.] // arXiv preprint. – 2022. – arXiv:2201.12328.

3. De, S. Unlocking High-Accuracy Differentially Private Image Classification through Scale / S. De, L. Berrada, J. Hayes [et al.] // arXiv preprint. – 2022. – arXiv:2204.13650.

4. Bu, Z. Deep Learning with Gaussian Differential Privacy / Z. Bu, J. Dong, Q. Long, W. J. Su // Harvard Data Science Review. – 2020. – Vol. 2, № 3. – P. 1–30. – DOI 10.1162/99608f92.cfc5dd25.

5. Ponomareva, N. How to DP-fy ML: A Practical Guide to Machine Learning with Differential Privacy / N. Ponomareva, J. Hazimeh, B. Kurakin [et al.] // Journal of Artificial Intelligence Research. – 2023. – Vol. 77. – P. 1113–1201. – DOI 10.1613/jair.1.14649.