УДК 004.89

МЕТОДЫ ПРОВЕРКИ УСТОЙЧИВОСТИ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ К ЦЕЛЕВЫМ ВОЗДЕЙСТВИЯМ

Калинюк А.О., Старший оператор научной роты, III курс
Военная академия связи имени Маршала Советского Союза С. М. Буденного,
г. Санкт-Петербург

Одним из ключевых вызовов, с которым сталкивается современное машинное обучение, является уязвимость моделей к различным формам внешнего целенаправленного вмешательства. Модель, продемонстрировавшая высокую точность на тестовых данных, может оказаться крайне нестабильной в условиях направленных атак – когда входные данные модифицируются таким образом, чтобы добиться конкретного неправильного вывода, при этом сохраняя внешнюю правдоподобность. Эти воздействия называются целевыми (targeted) атаками, и устойчивость модели к ним становится неотъемлемым требованием в критически значимых сферах применения, включая биометрию, судебную аналитику, автономные системы и кибербезопасность. Данная статья посвящена систематическому анализу методов проверки устойчивости моделей к целевым воздействиям, включая теоретические основания, практические техники, инфраструктурные решения и вызовы интерпретации [1].

Целевые атаки отличаются от ненаправленных тем, что в них атакующий стремится не просто вызвать ошибку модели, но добиться определённого желаемого предсказания. Например, злоумышленник может модифицировать изображение так, чтобы система распознавания лиц классифицировала его как конкретного пользователя, или сформировать текст, который будет ошибочно принят фильтром как «нейтральный», несмотря на скрытую токсичность. Учитывая потенциальную направленность атак на конкретные действия модели, методы оценки устойчивости к целевым воздействиям должны быть более тонкими и специализированными, чем стандартные процедуры тестирования на обобщающую способность [2].

Существуют различные подходы к проверке устойчивости модели в условиях целевых атак. Один из наиболее развитых – это градиентные атаки с заданным целевым классом. В их основе лежит модификация входа таким образом, чтобы модель максимально приблизилась к уверенной классификации в пользу заданной категории. Метод C&W (Carlini & Wagner), а также Targeted PGD (Projected Gradient Descent) и Targeted DeepFool являются примерами таких алгоритмов. Они строят оптимизационную задачу, где минимизируется

расстояние между модифицированным и исходным входом при одновременном максимальном усилении активации нужного целевого класса.

Проверка устойчивости заключается в оценке минимальной величины изменения, необходимого для переноса предсказания модели из истинного в заданный класс, что позволяет построить метрику направленной робастности.

Другим направлением являются атаки на текстовые и языковые модели, где целевые воздействия могут быть реализованы через переформулировку, замену слов с сохранением семантики или вставку специальных последовательностей (trigger words). Методы вроде TextFooler, BERT-Attack, SemAttack и PSO-based атак позволяют формировать примеры, в которых LLM или классификатор ошибочно интерпретирует вредоносный или некорректный контент как приемлемый. Проверка устойчивости включает анализ семантической стабильности модели, а также выявление слабых мест в области токенизации, внимания или механизма генерации. Особенно важны такие тесты в задачах модерации, юридической оценки и медицинской консультации, где ошибка может привести к социально значимым последствиям.

Особый интерес представляют физически реализуемые целевые атаки – например, изменение внешнего вида объектов, чтобы система компьютерного зрения классифицировала их ошибочно. Эксперименты с QR-наклейками на знаках, скамейках или элементах одежды показывают, что целевые атаки могут быть реализованы в реальной среде. Проверка устойчивости в этом контексте требует построения физически верифицируемых симуляций: от применения 3D-рендеринга до воспроизведения атак в лабораторных условиях с реальными камерами, освещением и движением. Это предполагает, что устойчивость модели должна проверяться в условиях вариативности среды и сенсорных искажений, что выходит за рамки классического цифрового тестирования.

В задачах федеративного обучения и распределённых систем, целевые воздействия могут быть реализованы через отравление данных (targeted data poisoning), когда атакующий клиент внедряет образцы, предназначенные для переноса модели в заданное поведение. Такие атаки могут быть тонкими, незаметными и выявляемыми лишь при наличии специализированных тестов. Методы оценки устойчивости включают здесь проверку влияния отдельных клиентов на агрегированную модель, симуляции коалиционных атак и статистический анализ поведения модели на локальных распределениях. Существуют также алгоритмы «replay attacks», при которых атакующие узлы вносят в модель предопределённые шаблоны поведения, вызывающие целевые сбои в критических точках.

Для комплексной оценки устойчивости необходима инфраструктура тестирования, включающая генераторы атак, профили угроз, симуляторы среды

и системы логирования. Развиваются фреймворки, поддерживающие целевые атаки в автоматическом режиме: Adversarial Robustness Toolbox, Foolbox, TextAttack, IBM ART. Они позволяют задать целевое поведение, варьировать силу атаки, проводить серийные тесты и собирать метрики. Важным аспектом становится также визуализация результатов – представление векторов смещений, изменения активаций, доверительных вероятностей и зон неустойчивости модели [3].

Однако даже при наличии развитого инструментария встаёт проблема интерпретации результатов. Малейшее смещение может быть как естественным, так и атакующим – различить их трудно. Кроме того, проверка устойчивости в условиях ограниченного знания (например, black-box целевые атаки) требует моделирования вероятностного поведения атакующего, что усложняет процедуру оценки [4].

Здесь перспективным направлением становится статистическое тестирование под ограничениями, где для каждой задачи строится набор допустимых трансформаций, и измеряется вероятность вывода модели, отличного от эталонного. Подобные методы требуют интердисциплинарной проработки с участием специалистов в области статистики, кибербезопасности и нормативного права [5].

Существуют и формально-верифицируемые методы проверки устойчивости. Они позволяют задать математически доказуемые границы, в пределах которых поведение модели не изменится, независимо от формы атакующего воздействия. Такие подходы, как randomized smoothing, Lipschitz bounding, interval propagation и SAT/SMT-верификация, набирают популярность в высокорисковых приложениях. Несмотря на их вычислительную сложность, они позволяют не просто эмпирически проверить модель, а гарантировать её безопасность в строго определённых условиях.

В заключение следует подчеркнуть, что проверка устойчивости моделей к целевым воздействиям является неотъемлемой частью проектирования доверенных ИИ-систем. Это не факультативный этап, а обязательная компонента инженерии надёжности, особенно в тех сферах, где на карту поставлена человеческая жизнь, безопасность и права.

Развитие методологий, фреймворков и стандартов оценки устойчивости позволит не только лучше понимать поведение моделей, но и строить ИИ, которому можно доверять даже в условиях враждебной и нестабильной среды.

Список литературы:

1. Eykholt, K. Robust Physical-World Attacks on Deep Learning Visual Classification / K. Eykholt, I. Evtimov, E. Fernandes [et al.] // 2018 IEEE/CVF

Conference on Computer Vision and Pattern Recognition. – Salt Lake City : IEEE, 2018. – P. 1625-1634. – DOI 10.1109/CVPR.2018.00175.

2. Evtimov, I. Robust Physical-World Attacks on Machine Learning Models / I. Evtimov, K. Eykholt, E. Fernandes [et al.] // arXiv. – 2017. – arXiv:1707.08945.

3. Athalye, A. Synthesizing Robust Adversarial Examples / A. Athalye, L. Engstrom, A. Ilyas, K. Kwok // Proceedings of the 35th International Conference on Machine Learning. – 2018. – Vol. 80. – P. 284-293.

4. Engstrom, L. A Rotation and a Translation Suffice : Fooling CNNs with Simple Transformations / L. Engstrom, B. Tran, D. Tsipras [et al.] // arXiv. – 2017. – arXiv:1712.02779.

5. Hosseini, H. Semantic Adversarial Examples / H. Hosseini, B. Xiao, M. Jaiswal, R. Poovendran // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. – Salt Lake City : IEEE, 2018. – P. 1614-1619. – DOI 10.1109/CVPRW.2018.00206.