УДК 004.89

ПОСТРОЕНИЕ УНИВЕРСАЛЬНОГО ФРЕЙМВОРКА ДЛЯ ОЦЕНКИ УСТОЙЧИВОСТИ К АТАКАМ

Панчак В.П., Старший оператор научной роты, III курс
Военная академия связи имени Маршала Советского Союза С. М. Буденного,
г. Санкт-Петербург

В условиях стремительного развития технологий искусственного интеллекта и машинного обучения, а также их широкого внедрения в критически важные сферы экономики, медицины, безопасности и промышленности, вопрос обеспечения устойчивости моделей к разнообразным атакам приобретает ключевое значение.

Современные нейросетевые архитектуры, включая глубокие свёрточные сети, трансформеры и гибридные модели, демонстрируют впечатляющие результаты на стандартных датасетах, однако при столкновении с целенаправленными попытками нарушения их работы – так называемыми адверсариальными атаками, шумами, пертурбациями и иными формами вмешательства – их поведение может резко деградировать.

В таких условиях становится необходимым создание универсального фреймворка для оценки устойчивости моделей, способного охватить широкий спектр атак и обеспечить воспроизводимую, объективную и масштабируемую экспертизу [1].

Построение универсального фреймворка требует решения множества комплексных задач, начиная от стандартизации форматов ввода и вывода моделей до разработки гибкой системы генерации атакующих сценариев и метрик оценки. Одной из ключевых особенностей такого фреймворка должна стать модульная архитектура, позволяющая интегрировать разнообразные методы атак – от классических шумовых трансформаций и искажающих фильтров до сложных семантически осмысленных атак, ориентированных на конкретные задачи и типы данных. Важным аспектом является возможность расширения системы новыми алгоритмами атаки и защитными механизмами без необходимости переработки базовой инфраструктуры, что обеспечивает долгосрочную актуальность и адаптивность фреймворка [2].

С точки зрения функциональности, универсальный фреймворк должен поддерживать работу с различными типами данных и моделей: изображения, текст, аудио, видео, многомодальные данные, что требует продуманной системы абстракций и интерфейсов. Это позволяет применять единые принципы

тестирования в самых разных областях, обеспечивая сопоставимость результатов и возможность комплексного анализа поведения моделей в гетерогенных условиях. Для каждого типа данных должны быть разработаны специализированные модули атак и трансформаций, учитывающие особенности входной информации и уязвимости соответствующих архитектур [3].

Особое внимание в рамках фреймворка уделяется метрикам оценки устойчивости. Традиционные показатели качества модели, такие как accuracy, precision, recall, зачастую оказываются недостаточными для полноты картины. Поэтому во фреймворк интегрируются продвинутые метрики, оценивающие изменения распределения выходных вероятностей, стабильность внутреннего латентного пространства, расстояния в эмбеддинговом пространстве, а также показатели уверенности и согласованности модели при атаках. Эти метрики могут быть как глобальными, отражающими общую устойчивость, так и локальными – показывающими чувствительность к конкретным типам атак или отдельным входным элементам. Комплексное использование таких показателей даёт возможность формировать подробные отчёты и рекомендации по повышению робастности [4].

Важной составляющей фреймворка является поддержка автоматизации тестирования и интеграции с MLOps-пайплайнами. Это позволяет выполнять регулярные проверки устойчивости моделей на разных этапах жизненного цикла – от начального обучения и валидации до выпуска в продуктив и последующего мониторинга. Автоматизация снижает человеческий фактор и ошибки, повышает скорость и качество оценки, а также обеспечивает воспроизводимость и аудит результатов, что становится критически важным в контексте нормативных требований и стандартов безопасности ИИ. Интеграция с системами управления версиями моделей, платформами мониторинга и аналитики позволяет формировать циклы обратной связи для постоянного улучшения и адаптации моделей [5].

Техническая реализация универсального фреймворка включает использование современных параллельных вычислений и распределённых систем для масштабирования генерации атак и анализа большого объёма данных. Поддержка различных языков программирования и стандартов обеспечивает широкое применение и совместимость с уже существующими инструментами и библиотеками. Важным аспектом является также удобство интерфейса, позволяющего исследователям и инженерам быстро настраивать сценарии тестирования, просматривать результаты и экспортировать отчёты в удобном формате. Интерактивные визуализации, позволяющие анализировать динамику поведения моделей при различных атаках, существенно повышают качество интерпретации и принятия решений.

Неотъемлемой частью фреймворка становится сообщество пользователей и разработчиков, обеспечивающее постоянный обмен новыми методами, атаками и защитами. Открытый исходный код, модульность и стандартизированные API способствуют быстрой адаптации к новым вызовам и быстрому внедрению инноваций. Это также стимулирует создание обширных бенчмарков устойчивости, сравнение моделей и развитие лучших практик в области безопасного и надёжного машинного обучения.

В итоге, построение универсального фреймворка для оценки устойчивости к атакам является фундаментальной задачей для повышения доверия к искусственному интеллекту. Такой фреймворк не только позволяет выявлять уязвимости и слабые места моделей, но и служит платформой для тестирования защитных механизмов, разработки более робастных архитектур и повышения общей безопасности ИИ-систем.

Он способствует формированию единого стандарта оценки устойчивости, упрощает процессы сертификации и аудита, а также обеспечивает прозрачность и воспроизводимость экспериментов. В условиях динамично меняющейся технологической и нормативной среды универсальный фреймворк становится незаменимым инструментом для исследователей, разработчиков и конечных пользователей, стремящихся создавать и эксплуатировать надёжные, эффективные и этически безопасные модели машинного обучения [6].

Список литературы:

1. Raghunathan, A. Certified Defenses against Adversarial Examples / A. Raghunathan, J. Steinhardt, P. Liang // International Conference on Learning Representations. – 2018.

2. Hein, M. Formal Guarantees on the Robustness of a Classifier against Adversarial Manipulation / M. Hein, M. Andriushchenko // Advances in Neural Information Processing Systems. – 2017. – Vol. 30.

3. Mirman, M. Differentiable Abstract Interpretation for Provably Robust Neural Networks / M. Mirman, T. Gehr, M. Vechev // Proceedings of the 35th International Conference on Machine Learning. – 2018. – Vol. 80. – P. 3578-3586.

4. Dvijotham, K. A Dual Approach to Scalable Verification of Deep Networks / K. Dvijotham, R. Stanforth, S. Gowal [et al.] // Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence. – 2018. – P. 550-559.

5. Raghunathan, A. Semidefinite Relaxations for Certifying Robustness to Adversarial Examples / A. Raghunathan, J. Steinhardt, P. Liang // Advances in Neural Information Processing Systems. – 2018. – Vol. 31.

6. Singh, G. Fast and Effective Robustness Certification / G. Singh, T. Gehr, M. Puschel, M. Vechev // Advances in Neural Information Processing Systems. – 2018. – Vol. 31.