УДК 004.89

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ИНСТРУМЕНТОВ ОЦЕНКИ УСТОЙЧИВОСТИ ОБУЧЕННЫХ МОДЕЛЕЙ

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

В последние годы внимание к проблеме устойчивости обученных моделей машинного обучения выросло не только в академических кругах, но и на уровне промышленного применения, регулирования и политических дискуссий. Причина этому – выявление масштабных уязвимостей даже в моделях, демонстрирующих высокую точность в лабораторных условиях. От малозаметных адверсариальных примеров до систематических ошибок при выходе за пределы обучающего распределения – всё это поднимает вопрос: насколько можно доверять результатам ИИ, особенно в критически чувствительных областях? Для ответа на этот вопрос требуется не столько единичный тест, сколько комплексная инфраструктура инструментов, способных оценивать устойчивость обученных моделей по множеству параметров. Настоящая работа направлена на сравнительный анализ таких инструментов, с выделением их архитектурных особенностей, поддерживаемых типов моделей, доступных атак, метрик и применимости в реальных пайплайнах разработки [1].

Под устойчивостью (robustness) модели в широком смысле понимается её способность сохранять предсказательную корректность при внесении изменений в входные данные, конфигурацию среды или параметры системы. Эти изменения могут быть случайными (например, шум), систематическими (например, смещение распределения), вредоносными (адверсариальные атаки), физически реализуемыми (реальные изменения в изображении), структурными (ошибки в графе), или даже социальными (влияние предвзятости в данных). Вследствие такой широты понятий, инструменты оценки устойчивости различаются по ориентации: одни ориентированы на визуальные модели, другие на текст, третьи – на табличные данные, четвёртые – на модели в реальном времени [2].

В рамках анализа были выбраны шесть наиболее зрелых и часто используемых фреймворков: CleverHans, IBM Adversarial Robustness Toolbox (ART), Foolbox, DeepSec, TextAttack и RobustBench. Каждый из них представляет собой открытый инструмент с активным сообществом, поддержкой до-

кументации и интеграцией в экосистему популярных ML-фреймворков (PyTorch, TensorFlow, Keras, HuggingFace и др.).

1. CleverHans

Это один из первых фреймворков для оценки устойчивости, разработанный в Google Brain. Он ориентирован на классические ad-hoc атаки (FGSM, JSMA, BIM, DeepFool) и интегрирован в TensorFlow. Несмотря на относительно устаревшую архитектуру и ограниченную поддержку новых методов, CleverHans популярен в учебных целях и при тестировании базовых устойчивостей. Основное ограничение – слабая поддержка современных методов сертификации и отсутствие нативной работы с PyTorch [3].

Преимущества: зрелость, документация, простота использования.Недостатки: неактуальные атаки, плохая поддержка новых фреймворков, нет текстовых моделей.

2. IBM Adversarial Robustness Toolbox (ART)

Это наиболее всеобъемлющий фреймворк, поддерживающий как white-box, так и black-box атаки, тестирование устойчивости, методы защиты, генерацию отчётов и интеграцию с промышленными пайплайнами. ART поддерживает обширный список атак (более 50), сертифицируемые методы, анализ приватности, интерпретируемость, защитные трансформации. Он также совместим с ONNX, что облегчает промышленную интеграцию.

Преимущества: широчайшая поддержка методов, текст, изображение, таблицы, документация от индустрии, поддержка сертификаций.Недостатки: высокая сложность настройки, громоздкость, возможная избыточность для простых сценариев.

3. Foolbox [4].

Foolbox ориентирован на модульное построение атак и глубокую поддержку PyTorch. Он позволяет строить кастомные атаки с использованием автодифференцирования и интеграции с Eager Execution. Также содержит интерфейсы для оценки сертифицированной устойчивости (например, с помощью randomized smoothing). Особенностью Foolbox является его ориентация на исследовательскую воспроизводимость: все атаки снабжены стандартными параметрами, что упрощает репликацию экспериментов [5].

Преимущества: гибкость, модульность, активное развитие, хорошие интерфейсы PyTorch.Недостатки: слабая поддержка табличных и текстовых моделей, нет стандартного GUI или отчётов.

4. DeepSec

Это фреймворк с ориентацией на строгость и воспроизводимость оценки, использующий единую экспериментальную парадигму. Поддерживает реализацию более 30 атак, множество защит и метрик (accuracy under attack, minimal perturbation, success rate и др.). Отличительной чертой DeepSec явля-

ется независимая иерархия атак/защит/метрик, а также автоматическое построение отчётов. Часто используется в академических исследованиях по сертифицированной робастности.

Преимущества: исследовательская строгость, воспроизводимость, удобство батчирования экспериментов.Недостатки: ориентирован на исследовательскую среду, неинтуитивный API для новичков.

5. TextAttack

Фреймворк, ориентированный на текстовые модели, включая классификацию, NLI, QA, генерацию. Поддерживает семантически осмысленные атаки, включая синтаксические переформулировки, замену синонимов, перестановки слов и другие лингвистически мотивированные стратегии. Также позволяет тренировать защищённые модели, использовать стандартные датасеты и генерировать визуализации. Это ведущий инструмент в области NLP-робастности [6].

Преимущества: разнообразие атак для текста, удобный интерфейс, визуализация, генерация adversarial datasets.Недостатки: слабая поддержка мультимодальных задач, высокая зависимость от HuggingFace.

6. RobustBench

Это бенчмарк-платформа для сравнения устойчивых моделей на общедоступных датасетах (CIFAR-10, ImageNet, TinyImageNet). Поддерживает только ограниченный набор атак, но делает акцент на стандартизированной оценке и ранжировании моделей, прошедших тестирование по фиксированным метрикам. Является эталоном для сертифицированных методов и защищённых архитектур.

Преимущества: прозрачность, повторяемость, стандартизированные метрики.

Недостатки: ограниченное число атак, не подходит для кастомного анализа.

Сравнительный анализ показывает, что выбор инструмента зависит от целей и контекста применения. Для промышленного внедрения и нормативной отчётности предпочтителен ART, как наиболее зрелый и комплексный. Для академических исследований устойчивости изображений подходят DeepSec и Foolbox. Для NLP-приложений оптимален TextAttack. А для публикаций и сравнения моделей незаменимым является RobustBench. Развитие таких инструментов создаёт основу для формирования культуры тестирования моделей не только на точность, но и на надёжность, безопасность, объяснимость – то есть доверие в инженерии ИИ, что становится ключевым вызовом новой технологической эпохи.

Список литературы:

1. Croce, F. Reliable Evaluation of Adversarial Robustness with an Ensemble of Diverse Parameter-free Attacks / F. Croce, M. Hein // Proceedings of the 37th International Conference on Machine Learning. – 2020. – Vol. 119. – P. 2206-2216.

2. Andriushchenko, M. Square Attack : A Query-Efficient Black-Box Adversarial Attack via Random Search / M. Andriushchenko, F. Croce, N. Flammarion, M. Hein // Computer Vision - ECCV 2020. – Cham : Springer, 2020. – P. 484-501. – DOI 10.1007/978-3-030-58592-1_29.

3. Brendel, W. Decision-Based Adversarial Attacks : Reliable Attacks against Black-Box Machine Learning Models / W. Brendel, J. Rauber, M. Bethge // International Conference on Learning Representations. – 2018.

4. Rauber, J. Foolbox : A Python Toolbox to Benchmark the Robustness of Machine Learning Models / J. Rauber, W. Brendel, M. Bethge // Reliable Machine Learning in the Wild Workshop, International Conference on Machine Learning. – 2017.

5. Modas, A. SparseFool : A Few Pixels Make a Big Difference / A. Modas, S.-M. Moosavi-Dezfooli, P. Frossard // 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. – Long Beach : IEEE, 2019. – P. 9087-9096. – DOI 10.1109/CVPR.2019.00930.

6. Laidlaw, C. Functional Adversarial Attacks / C. Laidlaw, S. Feizi // Advances in Neural Information Processing Systems. – 2019. – Vol. 32.