УДК 004.89

ОЦЕНКА УСТОЙЧИВОСТИ НЕЙРОСЕТЕЙ С ИСПОЛЬЗОВАНИЕМ СИНТЕТИЧЕСКИХ АТАК

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

С увеличением распространения нейросетевых моделей в критически важных приложениях – от медицинской диагностики и автономного транспорта до финансового анализа и судебных решений – на первый план выходит задача оценки их устойчивости к внешним вмешательствам. Нейросети, несмотря на свою мощность, уязвимы к целому классу воздействий, известных как адверсариальные атаки. Это такие преднамеренные изменения входных данных, которые остаются практически незаметными для человека, но способны радикально изменить поведение модели. Для надёжной эксплуатации ИИ-систем в условиях открытого и потенциально враждебного мира необходимо вырабатывать чёткие процедуры оценки устойчивости моделей, и одним из наиболее перспективных подходов является использование синтетических атак – то есть специально сконструированных сценариев, моделирующих целенаправленные и разнообразные угрозы. В данной статье рассматриваются принципы, методы и вызовы, связанные с таким подходом [1].

Оценка устойчивости моделей машинного обучения должна быть системной, воспроизводимой и охватывать как тривиальные, так и изощрённые сценарии вторжений. При этом традиционные методы тестирования, основанные на эмпирической проверке модели на отложенных данных, не выявляют всех потенциальных уязвимостей, поскольку такие данные не включают в себя преднамеренно вредоносные примеры. Именно по этой причине исследовательское сообщество всё активнее разрабатывает искусственные сценарии тестирования, при которых входные данные модифицируются согласно определённой стратегии атаки. Такие атаки могут моделировать искажённую подачу данных в реальной среде, ошибки сбора, злонамеренные вмешательства или ошибки интерпретации модели.

Наиболее известной формой синтетических атак являются градиентные адверсариальные атаки. Их суть заключается в том, что градиент целевой функции модели по входным данным используется для генерации направленного, но минимального изменения входа. Один из классических методов – FGSM (Fast Gradient Sign Method), при котором к входному вектору прибавля-

ется вектор, направленный по знаку градиента. Более точные методы – PGD (Projected Gradient Descent), DeepFool, C&W attack – используют итерационные схемы и оптимизационные процедуры для генерации атак, нацеленных на точечный вывод модели. Эти методы позволяют точно оценить, насколько легко можно обмануть модель, используя только информацию о её структуре и градиентах (в white-box-сценарии).

Кроме градиентных атак, существуют и black-box атаки, при которых модель рассматривается как «чёрный ящик», и доступ к её внутренним параметрам отсутствует. В этом случае используются эвристические методы, такие как атакующие генеративные модели, эволюционные алгоритмы, методики замены моделей (substitute model attacks) и аппроксимации. Особую роль играют трансферные атаки, при которых примеры, созданные для одной модели, успешно применяются к другой, схожей по архитектуре. Такие сценарии имеют высокую практическую значимость, так как в реальном мире злоумышленник редко обладает полным доступом к модели.

Синтетические атаки могут быть нацелены не только на классификацию, но и на другие задачи: сегментацию изображений, машинный перевод, генерацию текста, извлечение признаков, рекомендации. Кроме того, атаки могут быть направленными (targeted) и ненаправленными (untargeted), в зависимости от того, пытается ли атакующий добиться конкретного ложного результата или любого неправильного результата. Важно также учитывать физическую реалистичность атак – например, возможность реализовать их через печать изображений, добавление наклеек, изменение акустических сигналов и т.п., что актуализирует использование физически реализуемых синтетических атак.

Практическое применение синтетических атак в оценке устойчивости нейросетей требует целой инфраструктуры: набора генераторов атак, тестовых сценариев, автоматизированных отчётов и инструментов визуализации. В последние годы появились фреймворки, такие как CleverHans, Foolbox, Adversarial Robustness Toolbox (ART), TextAttack, DeepSec, которые позволяют встраивать процедуры атак в пайплайны обучения и тестирования. Их использование позволяет систематизировать процесс: применять множество атак по шаблону, рассчитывать метрики (например, accuracy under attack, certified radius, robustness score), анализировать влияние архитектурных изменений, методов защиты и предварительной обработки [2].

Важнейшим аспектом является интерпретация результатов атак. Простое снижение точности на 5% в условиях атаки ещё не говорит о катастрофической уязвимости, тогда как резкое изменение предсказаний при минимальных воздействиях – тревожный сигнал. Поэтому всё большее внимание уделяется устойчивым метрикам, таким как средняя минимальная величина

возмущения, необходимая для смены предсказания, или доверительные интервалы вероятностного вывода под атакой. Существуют также сертифицируемые методы устойчивости, при которых гарантируется, что модель не изменит своё поведение в определённой области входного пространства (например, randomized smoothing, interval bound propagation), что становится особенно важно для высоконадежных приложений [3].

Однако стоит отметить, что оценка устойчивости с помощью синтетических атак – это лишь одна сторона проблемы. Важно сочетать этот подход с другими методами: анализом доверительных интервалов, тестированием на изменённых данных, исследованием out-of-distribution поведения, симуляциями пользовательских взаимодействий. Только такой комплексный подход позволяет дать более полную оценку реальной надёжности системы в условиях неопределённости и потенциального противодействия [4].

В заключение можно отметить, что оценка устойчивости нейросетей с использованием синтетических атакстановится неотъемлемой частью современного цикла разработки доверенных ИИ-систем. Она позволяет выявлять скрытые уязвимости, сравнивать архитектуры, проверять эффективность защит и формировать модели, устойчивые к реальным угрозам. В дальнейшем такие атаки будут всё чаще использоваться не только в академических исследованиях, но и в нормативных процедурах сертификации ИИ, в системах мониторинга поведения моделей на продакшене, в построении самоконтролируемых ИИ-агентов. Развитие стандартизированных фреймворков, баз сценариев атак и открытых отчётов станет важнейшим условием построения надёжной и этически устойчивой инфраструктуры искусственного интеллекта [5, 6].

Список литературы:

1. Moosavi-Dezfooli, S.-M. DeepFool : A Simple and Accurate Method to Fool Deep Neural Networks / S.-M. Moosavi-Dezfooli, A. Fawzi, P. Frossard // 2016 IEEE Conference on Computer Vision and Pattern Recognition. – Las Vegas : IEEE, 2016. – P. 2574-2582. – DOI 10.1109/CVPR.2016.282.

2. Carlini, N. Towards Evaluating the Robustness of Neural Networks / N. Carlini, D. Wagner // 2017 IEEE Symposium on Security and Privacy. – San Jose : IEEE, 2017. – P. 39-57. – DOI 10.1109/SP.2017.49.

3. Papernot, N. Practical Black-Box Attacks against Machine Learning / N. Papernot, P. McDaniel, I. Goodfellow [et al.] // Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security. – New York : ACM, 2017. – P. 506-519. – DOI 10.1145/3052973.3053009.

4. Kurakin, A. Adversarial Examples in the Physical World / A. Kurakin, I. Goodfellow, S. Bengio // International Conference on Learning Representations Workshop. – 2017.

5. Brown, T. B. Adversarial Patch / T. B. Brown, D. Mane, A. Roy, M. Abadi, J. Gilmer // arXiv. – 2017. – arXiv:1712.09665.

6. Chen, P.-Y. Zoo : Zeroth Order Optimization Based Black-box Attacks to Deep Neural Networks without Training Substitute Models / P.-Y. Chen, H. Zhang, Y. Sharma [et al.] // Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. – New York : ACM, 2017. – P. 15-26. – DOI 10.1145/3128572.3140448.