УДК 004.89

ОЦЕНКА УСТОЙЧИВОСТИ МОДЕЛЕЙ К СЛУЧАЙНЫМ И НАПРАВЛЕННЫМ ИСКАЖЕНИЯМ

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

В эпоху стремительного роста применения машинного обучения в самых разных сферах человеческой деятельности, от медицины и финансов до автономных транспортных систем и безопасности, вопрос устойчивости моделей к различным видам искажений приобретает особую значимость.

Модели машинного обучения, особенно глубокие нейросети, демонстрируют высокую точность на тестовых данных, однако реальная эксплуатация зачастую сопряжена с воздействием не только случайных, но и целенаправленных искажений, которые могут существенно повлиять на качество и надёжность их работы. В связи с этим систематическая и комплексная оценка устойчивости моделей к этим видам искажений является одним из ключевых направлений современных исследований и разработок в области доверенного искусственного интеллекта.

Оценка устойчивости моделей к случайным искажениям связана с анализом поведения алгоритмов в условиях естественного шума и непреднамеренных ошибок, которые неизбежно возникают при сборе, передаче и обработке данных. Эти искажения могут принимать самые разные формы: от помех в сенсорных данных и шумов изображения до пропусков информации и артефактов, появляющихся в результате технических сбоев [1].

Случайные искажения представляют собой непредсказуемые изменения входных данных, не имеющие зловредного характера, однако они способны ухудшить качество предсказаний и вызвать нестабильность работы модели. Эффективная оценка устойчивости к таким воздействиям предполагает создание реалистичных сценариев тестирования, которые отражают реальные условия эксплуатации и позволяют выявлять критические пороги, за которыми модель начинает демонстрировать существенные ошибки [2].

В то же время направленные искажения, часто именуемые адверсариальными атаками, представляют собой специально спроектированные изменения входных данных, цель которых — вызвать ошибочное поведение модели или манипулировать её выводами. В отличие от случайных искажений, направленные атаки обладают чёткой стратегией и оптимизированы для обхо-

да защитных механизмов. Такие воздействия могут быть минимальными по амплитуде, но максимальными по своему эффекту, что делает их особенно опасными в реальных сценариях, связанных с безопасностью и защитой информации. Оценка устойчивости к направленным искажениям требует разработки специализированных методов генерации атакующих примеров, которые учитывают особенности архитектуры модели и задачи, а также проведение детального анализа реакций системы на подобные вмешательства [3].

Для комплексной оценки устойчивости необходима интеграция методов, позволяющих тестировать модели как на случайные, так и на направленные искажения в едином фреймворке. Такой подход даёт возможность сравнивать уязвимости модели, выявлять общие закономерности и особенности поведения, а также разрабатывать универсальные механизмы защиты. Практически это реализуется через генерацию разнообразных сценариев тестирования, включающих широкий спектр трансформаций данных: шумы различных типов и спектров, искажения геометрии, изменения яркости и контрастности, а также продвинутые адверсариальные методы — FGSM, PGD, CW и другие. Важной составляющей становится автоматизация процесса тестирования, позволяющая проводить массовые испытания с последующим анализом результатов.

Метрики оценки устойчивости должны выходить за рамки традиционных показателей точности, поскольку влияние искажений часто проявляется в снижении уверенности модели, изменении распределения вероятностей и росте непредсказуемых ошибок. Современные исследования предлагают использовать разнообразные показатели, отражающие стабильность предсказаний, расстояния между распределениями выходных значений, а также метрики, основанные на анализе внутренних представлений модели. Такой многоуровневый подход позволяет выявлять слабые места и обеспечивать более глубокое понимание механизмов возникновения нестабильности.

Особое внимание уделяется тестированию в условиях, максимально приближённых к реальным, включая динамические среды, мультисенсорные данные и сценарии, где искажения могут сочетаться или сменять друг друга во времени. Это позволяет оценивать не только реакцию на отдельные воздействия, но и устойчивость модели к комплексным и последовательным атакам, что имеет критическое значение для систем с долгим циклом работы и сложной архитектурой. Анализ таких сценариев помогает выявлять особенности адаптивного поведения моделей и возможности восстановления после сбоев.

Разработка и внедрение эффективных методик оценки устойчивости к случайным и направленным искажениям стимулирует создание новых архитектур и алгоритмов с повышенной робастностью. Среди них — использова-

ние методов регуляризации, аугментации данных, ансамблей моделей, а также специализированных слоёв и блоков, устойчивых к шуму и атакам [4].

Кроме того, полученные в процессе тестирования знания служат основой для разработки превентивных механизмов обнаружения и нейтрализации атакующих воздействий [5].

В конечном итоге, всесторонняя оценка устойчивости моделей к случайным и направленным искажениям является фундаментальной составляющей процесса создания доверенных систем машинного обучения. Она не только способствует выявлению и устранению уязвимостей, но и формирует базу для сертификации, стандартизации и регулирования применения ИИ в критически важных областях [6].

Реализация таких методик позволяет повысить безопасность, надёжность и прозрачность интеллектуальных систем, что существенно расширяет возможности их практического применения и способствует укреплению доверия пользователей и общества в целом.

Список литературы:

1. Dodge, S. Understanding How Image Quality Affects Deep Neural Networks / S. Dodge, L. Karam // 2016 Eighth International Conference on Quality of Multimedia Experience. – Lisbon : IEEE, 2016. – P. 1-6. – DOI 10.1109/QoMEX.2016.7498955.

2. Geirhos, R. ImageNet-trained CNNs Are Biased towards Texture; Increasing Shape Bias Improves Accuracy and Robustness / R. Geirhos, P. Rubisch, C. Michaelis [et al.] // International Conference on Learning Representations. – 2019.

3. Hendrycks, D. AugMix : A Simple Data Processing Method to Improve Robustness and Uncertainty / D. Hendrycks, N. Mu, E. D. Cubuk [et al.] // International Conference on Learning Representations. – 2020.

4. Yun, S. CutMix : Regularization Strategy to Train Strong Classifiers with Localizable Features / S. Yun, D. Han, S. J. Oh [et al.] // 2019 IEEE/CVF International Conference on Computer Vision. – Seoul : IEEE, 2019. – P. 6023-6032. – DOI 10.1109/ICCV.2019.00612.

5. Zhang, H. Mixup : Beyond Empirical Risk Minimization / H. Zhang, M. Cisse, Y. N. Dauphin, D. Lopez-Paz // International Conference on Learning Representations. – 2018.

6. Yin, D. A Fourier Perspective on Model Robustness in Computer Vision / D. Yin, R. Gontijo Lopes, J. Shlens, E. Cubuk, J. Gilmer // Advances in Neural Information Processing Systems. – 2019. – Vol. 32.