УДК 004.89

МЕТОДИКА ОЦЕНКИ УСТОЙЧИВОСТИ ОБУЧЕННЫХ МОДЕЛЕЙ В РЕАЛЬНЫХ СЦЕНАРИЯХ

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

В современных условиях стремительного распространения систем искусственного интеллекта и машинного обучения особенно остро встаёт проблема оценки устойчивости обученных моделей в реальных сценариях эксплуатации [1].

Теоретические показатели точности и другие стандартные метрики зачастую не отражают истинного поведения модели при воздействии разнообразных факторов, присущих реальной среде, включая шумы, искажения данных, непредвиденные ситуации и взаимодействия с человеческим фактором. Поэтому разработка методик, способных адекватно оценивать устойчивость моделей в условиях реальной эксплуатации, становится ключевым элементом обеспечения надёжности, безопасности и доверия к искусственному интеллекту.

Оценка устойчивости в реальных сценариях требует комплексного подхода, включающего моделирование и анализ широкого спектра возможных воздействий на модель. В отличие от лабораторных условий, где данные тщательно подготовлены и контролируются, реальная среда характеризуется множеством непредсказуемых и часто неструктурированных факторов. К таким факторам относятся динамические изменения во входных данных, вариативность среды, появление новых классов или аномалий, а также сложные взаимосвязи между элементами системы [2].

В результате поведение модели становится менее предсказуемым, что может привести к серьёзным ошибкам, включая аварийные ситуации в критических приложениях [3].

Для адекватной оценки устойчивости разработана методика, которая основывается на интеграции имитационного моделирования, сбора данных из производственной среды и анализа реакций модели на реальные и синтетические воздействия. Первый этап методики включает в себя построение разнообразных сценариев, максимально приближённых к условиям эксплуатации. Эти сценарии могут охватывать вариации качества и полноты данных, временные задержки, сбои в сенсорах и изменения окружающей среды.

Важным аспектом является использование синтетических атак и трансформаций, имитирующих реальные и потенциальные угрозы, включая адверсариальные воздействия, шумы, пропуски данных и др.

Второй этап предполагает систематический сбор статистики работы модели в реальных условиях, с фиксацией параметров среды, состояния входных данных и выходов модели. Такой мониторинг позволяет выявлять закономерности деградации качества, появление ошибок и зоны нестабильности [4].

Особое внимание уделяется анализу случаев, когда модель демонстрирует отклонения от ожидаемого поведения, что может свидетельствовать о необходимости дообучения, корректировки архитектуры или изменения подходов к сбору данных.

Третий этап методики предусматривает применение специализированных метрик, способных количественно оценить устойчивость в условиях вариативности и неопределённости.

Помимо традиционных показателей, таких как accuracy, precision, recall и F1-score, используются метрики, отражающие степень изменения выходных значений при воздействии трансформаций, уверенность модели, устойчивость распределения вероятностей и предсказательную стабильность. Важным инструментом становятся показатели, основанные на сравнении латентных представлений модели до и после воздействия, что позволяет глубже понять механизмы возникновения нестабильности.

Интеграция данной методики в жизненный цикл разработки и эксплуатации моделей предполагает использование автоматизированных платформ мониторинга и тестирования, которые позволяют в режиме реального времени отслеживать параметры устойчивости, выявлять тренды и аномалии, а также быстро реагировать на изменения среды.

Такой подход способствует формированию замкнутого цикла контроля качества – от обучения и валидации модели до её адаптации и обновления на основе полученных данных.

Заключение методики направлено на формализацию процесса оценки устойчивости, что существенно повышает прозрачность и воспроизводимость экспериментов, а также упрощает коммуникацию между разработчиками, исследователями и конечными пользователями. В условиях роста нормативных требований и стандартов по безопасности ИИ это становится важным фактором доверия и ответственности.

В итоге предложенная методика позволяет не только выявлять слабые места и уязвимости моделей, но и создавать условия для их своевременного исправления и адаптации, обеспечивая надёжность и безопасность интеллектуальных систем в реальных условиях эксплуатации [5, 6].

Список литературы:

1. Koh, P. W. WILDS : A Benchmark of in-the-Wild Distribution Shifts / P. W. Koh, S. Sagawa, H. Marklund [et al.] // Proceedings of the 38th International Conference on Machine Learning. – 2021. – Vol. 139. – P. 5637-5664.

2. Taori, R. Measuring Robustness to Natural Distribution Shifts in Image Classification / R. Taori, A. Dave, V. Shankar [et al.] // Advances in Neural Information Processing Systems. – 2020. – Vol. 33. – P. 18583-18599.

3. Sagawa, S. Distributionally Robust Neural Networks for Group Shifts : On the Importance of Regularization for Worst-Case Generalization / S. Sagawa, P. W. Koh, T. B. Hashimoto, P. Liang // International Conference on Learning Representations. – 2020.

4. Hendrycks, D. Natural Adversarial Examples / D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, D. Song // 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. – Nashville : IEEE, 2021. – P. 15262-15271. – DOI 10.1109/CVPR46437.2021.01502.

5. Miller, J. C. Accuracy on the Line : On the Strong Correlation between Out-of-Distribution and In-Distribution Generalization / J. C. Miller, A. Taori, A. Raghunathan [et al.] // Proceedings of the 38th International Conference on Machine Learning. – 2021. – Vol. 139. – P. 7721-7735.

6. Santurkar, S. How Does Batch Normalization Help Optimization? / S. Santurkar, D. Tsipras, A. Ilyas, A. Madry // Advances in Neural Information Processing Systems. – 2018. – Vol. 31.