УДК 004.89

ИНСТРУМЕНТЫ КОНТРОЛЯ КАЧЕСТВА УСТОЙЧИВОСТИ МОДЕЛЕЙ В ПРИКЛАДНЫХ ЗАДАЧАХ

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

В современном развитии технологий искусственного интеллекта и машинного обучения особое значение приобретает не только создание высокоточных моделей, но и обеспечение их устойчивости и надёжности в реальных прикладных задачах. В условиях динамично меняющейся среды эксплуатации, ограниченности данных, разнообразия источников шума и потенциальных атак устойчивость модели становится критическим параметром, напрямую влияющим на безопасность, качество и эффективность систем. В связи с этим разработка и внедрение инструментов контроля качества устойчивости моделей в прикладных задачах представляет собой одну из ключевых областей исследований и практических разработок, направленных на повышение доверия к интеллектуальным системам и обеспечение их стабильной работы [1].

Контроль качества устойчивости предполагает систематическую оценку поведения моделей при воздействии различных видов искажений и нестандартных ситуаций, возникающих в процессе эксплуатации. Для этого используются инструменты, способные моделировать и имитировать реальные сценарии, включающие как случайные шумы и помехи, так и целенаправленные атаки, направленные на нарушение корректности работы системы. Такие инструменты должны обеспечивать не только генерацию и внедрение воздействий, но и сбор, анализ и визуализацию результатов, позволяя выявлять слабые места модели, оценивать эффективность защитных механизмов и формировать рекомендации по их улучшению [2].

Одной из важных особенностей инструментов контроля качества устойчивости является их адаптивность и гибкость к специфике прикладных задач. Различные сферы применения — медицина, промышленность, финансы, транспорт — предъявляют свои требования к типам данных, формату входной информации и характеру потенциальных угроз. Поэтому эффективные инструменты должны поддерживать широкий спектр форматов и протоколов взаимодействия, обеспечивать возможность интеграции с существующими платформами машинного обучения и аналитики, а также предусматривать ка-

стомизацию сценариев тестирования в соответствии с особенностями конкретного применения [3].

Технически такие инструменты включают модули генерации атак и искажений, которые реализуют как классические методы шумового воздействия, так и современные адверсариальные атаки, методы аугментации и трансформации данных. Помимо этого, они оснащены средствами мониторинга производительности модели, сбора метрик устойчивости и анализа выходных данных, включая визуализацию изменений в распределениях, точности предсказаний, уверенности модели и стабильности внутренних представлений. Такой комплексный функционал обеспечивает глубокий и всесторонний анализ, необходимый для полноценного контроля качества [4].

Особое значение имеет автоматизация процессов контроля, позволяющая интегрировать инструменты в пайплайны MLOps и DevOps, что обеспечивает регулярный мониторинг устойчивости в реальном времени и своевременное реагирование на возникающие проблемы. Автоматизированные системы способны запускать тесты по расписанию или по событию, собирать и агрегировать данные, формировать отчёты и уведомления для разработчиков и аналитиков. Это значительно повышает эффективность контроля и снижает риски, связанные с человеческим фактором и задержками в обнаружении проблем.

Важным аспектом является также поддержка масштабируемости инструментов контроля качества устойчивости. Современные прикладные задачи часто связаны с обработкой больших объёмов данных и сложными моделями с миллионами параметров, что требует эффективного распределения вычислительных ресурсов и оптимизации процессов тестирования. Использование облачных платформ, контейнеризации и параллельных вычислений позволяет обеспечить необходимую производительность и гибкость, а также адаптироваться к росту объёмов данных и усложнению моделей без потери качества анализа [5].

Кроме технических возможностей, инструменты контроля качества устойчивости играют ключевую роль в обеспечении прозрачности и воспроизводимости экспериментов, что важно для научных исследований, промышленного внедрения и соблюдения нормативных требований. Наличие подробной документации, возможности сохранения и повторного запуска сценариев тестирования, а также интеграция с системами управления версиями и аудитом создают условия для системного подхода к оценке устойчивости и повышения доверия к моделям.

Наконец, развитие таких инструментов способствует формированию экосистемы доверенного машинного обучения, объединяющей разработчиков, исследователей и пользователей. Открытые стандарты, обмен знаниями и

совместная работа над улучшением методов контроля устойчивости создают предпосылки для быстрого реагирования на новые вызовы и угрозы, а также для внедрения инновационных подходов к обеспечению безопасности и надёжности ИИ-систем.

В итоге, инструменты контроля качества устойчивости моделей в прикладных задачах представляют собой необходимый элемент современной инфраструктуры машинного обучения, обеспечивая системный, комплексный и эффективный подход к выявлению, анализу и минимизации рисков, связанных с воздействием искажений и атак. Их развитие и внедрение позволяют создавать более надёжные, безопасные и адаптивные интеллектуальные системы, способные успешно функционировать в реальных, динамичных и зачастую неопределённых условиях эксплуатации.

Список литературы:

1. Amodei, D. Concrete Problems in AI Safety / D. Amodei, C. Olah, J. Steinhardt [et al.] // arXiv. – 2016. – arXiv:1606.06565.

2. Sambasivan, N. Re-imagining Algorithmic Fairness in India and Beyond / N. Sambasivan, E. Arnesen, B. Hutchinson, T. Doshi, V. Prabhakaran // Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. – New York : ACM, 2021. – P. 315-328. – DOI 10.1145/3442188.3445896.

3. Nushi, B. Towards Accountable AI : Hybrid Human-Machine Analyses for Characterizing System Failure / B. Nushi, E. Kamar, E. Horvitz // Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. – New York : ACM, 2018. – P. 126-132. – DOI 10.1145/3278721.3278778.

4. Kumar, A. Verified Uncertainty Calibration / A. Kumar, P. S. Liang, T. Ma // Advances in Neural Information Processing Systems. – 2019. – Vol. 32.

5. D’Amour, A. Underspecification Presents Challenges for Credibility in Modern Machine Learning / A. D’Amour, K. Heller, D. Moldovan [et al.] // Journal of Machine Learning Research. – 2022. – Vol. 23, № 226. – P. 1-61.