УДК 004.89

ФОРМАЛИЗАЦИЯ КРИТЕРИЕВ ОЦЕНКИ УСТОЙЧИВОСТИ ОБУЧЕННЫХ МОДЕЛЕЙ К ВНЕШНИМ ВОЗДЕЙСТВИЯМ

Мелихов И.А., Оператор научной роты, III курс
Военная академия связи имени Маршала Советского Союза С. М. Буденного,
г. Санкт-Петербург

В условиях быстрого развития искусственного интеллекта и его повсеместного внедрения во все сферы жизни особое внимание уделяется не только созданию высокоточных моделей, но и обеспечению их устойчивости к разнообразным внешним воздействиям. От качества этой устойчивости напрямую зависит надежность, безопасность и эффективность систем машинного обучения, особенно в критически важных приложениях – здравоохранении, финансах, промышленности и безопасности. Однако до сих пор одной из значимых проблем остается отсутствие единой формализованной базы для оценки устойчивости моделей, что затрудняет стандартизацию, сравнение и контроль качества ИИ-систем. Формализация критериев оценки устойчивости обученных моделей представляет собой фундаментальную задачу, необходимую для построения доверенных и надежных интеллектуальных систем [1].

Начнем с того, что устойчивость модели – это комплексное понятие, включающее способность сохранять качество работы и стабильность предсказаний при воздействии различных видов искажений, шумов, атак и других факторов, способных изменять входные данные или внутренние состояния. Для объективной оценки этой способности требуется разработка четко определенных критериев, которые позволят системно и однозначно измерять устойчивость, выявлять слабые места и формировать рекомендации по улучшению моделей. Формализация таких критериев подразумевает не просто перечисление качественных характеристик, но и создание математических моделей, метрик и процедур, способных отражать реальное поведение систем в условиях воздействия [2].

Одним из ключевых аспектов формализации является определение классов воздействий, к которым должна проявлять устойчивость модель. Эти воздействия могут включать случайные шумы, искажения, пертурбации, а также направленные адверсариальные атаки, вариации среды и динамические изменения данных. Формализация требует введения формальных описаний этих воздействий, их параметризации и классификации, что позволит создавать тестовые сценарии и проводить сопоставимый анализ. Кроме того, кри-

терии должны учитывать специфику прикладных задач и требований, что требует разработки адаптивных и контекстно-зависимых метрик.

Метрики устойчивости должны быть многоуровневыми и комплексными. Традиционные показатели качества, такие как accuracy или F1-score, не отражают полноты картины устойчивости, поэтому в формализацию включаются метрики, оценивающие чувствительность модели к изменениям входных данных, стабильность распределения выходов, устойчивость внутренних представлений и изменчивость вероятностных предсказаний. Важное место занимают вероятностные и статистические методы, позволяющие оценивать уровень неопределенности и вариабельности модели под воздействием возмущений. Формализованные критерии должны обеспечивать количественные границы допустимых изменений и служить основой для сравнительного анализа.

Для практической реализации формализации разрабатываются стандартизированные процедуры тестирования и наборы данных, позволяющие воспроизводимо проверять модели в контролируемых условиях. Это включает создание эталонных бенчмарков и фреймворков, в которых формализованные критерии внедрены в инструментарий оценки. Автоматизация и интеграция с современными MLOps процессами обеспечивает постоянный контроль качества устойчивости на всех этапах жизненного цикла модели – от обучения и валидации до эксплуатации и обновления.

Особое внимание уделяется возможности формализации в условиях ограниченного доступа к данным и необходимости защиты конфиденциальной информации. Критерии должны быть совместимы с методами федеративного обучения, дифференциальной приватности и другими технологиями защиты, что позволяет проводить оценку устойчивости без нарушения требований безопасности и приватности. Это расширяет возможности применения формализованных подходов в чувствительных и регламентированных областях.

Системный подход к формализации критериев способствует не только технической стандартизации, но и формированию общей терминологии и методологической базы, что важно для научного сообщества, промышленности и регуляторов. Единые формализованные критерии позволяют повысить прозрачность и воспроизводимость исследований, сравнивать различные модели и подходы, а также создавать нормативные документы и стандарты качества в области доверенного искусственного интеллекта [3].

В конечном счете, формализация критериев оценки устойчивости обученных моделей к внешним воздействиям является необходимым шагом на пути к созданию действительно надёжных, безопасных и адаптивных ИИ-систем. Это позволяет системно выявлять уязвимости, разрабатывать эффек-

тивные методы защиты и обеспечивать высокий уровень доверия со стороны пользователей и общества в целом. Продолжение исследований и внедрение формализованных подходов станет ключевым фактором устойчивого развития и успешного применения искусственного интеллекта в самых разных сферах жизни [4-6].

Список литературы:

1. Bubeck, S. Convex Optimization : Algorithms and Complexity / S. Bubeck // Foundations and Trends in Machine Learning. – 2015. – Vol. 8, № 3-4. – P. 231-357. – DOI 10.1561/2200000050.

2. Weng, T.-W. Evaluating the Robustness of Neural Networks : An Extreme Value Theory Approach / T.-W. Weng, H. Zhang, H. Chen [et al.] // International Conference on Learning Representations. – 2018.

3. Weng, T.-W. Towards Fast Computation of Certified Robustness for ReLU Networks / T.-W. Weng, H. Zhang, H. Chen [et al.] // Proceedings of the 35th International Conference on Machine Learning. – 2018. – Vol. 80. – P. 5276-5285.

4. Farnia, F. Generalizable Adversarial Training via Spectral Normalization / F. Farnia, J. Zhang, D. Tse // International Conference on Learning Representations. – 2019.

5. Lecuyer, M. Certified Robustness to Adversarial Examples with Differential Privacy / M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, S. Jana // 2019 IEEE Symposium on Security and Privacy. – San Francisco : IEEE, 2019. – P. 656-672. – DOI 10.1109/SP.2019.00044.

6. Salman, H. A Convex Relaxation Barrier to Tight Robustness Verification of Neural Networks / H. Salman, G. Yang, H. Zhang, C.-J. Hsieh, P. Zhang // Advances in Neural Information Processing Systems. – 2019. – Vol. 32.