УДК 004.89

ИНСТРУМЕНТЫ ВЕРИФИКАЦИИ И СЕРТИФИКАЦИИ ПОВЕДЕНИЯ ПРЕДОБУЧЕННЫХ МОДЕЛЕЙ

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

С масштабным внедрением предобученных моделей в критически важные сферы – от медицины и транспорта до финансов и государственной экспертизы – всё более актуальной становится необходимость формального подтверждения корректности и допустимости поведения таких моделей. Высокая производительность на тестовых выборках, даже сопровождаемая улучшенными метриками устойчивости или интерпретируемости, не является достаточной гарантией надёжности в условиях реального применения. Особенно это касается тех случаев, когда решения модели становятся основанием для действий, затрагивающих здоровье, свободу, безопасность или финансы человека.

В этих условиях требуется не просто техническое тестирование, а инструменты верификации и сертификации поведения предобученных моделей, способные обеспечить соответствие моделей требованиям нормативной, этической и функциональной надёжности [1].

Верификация (verification) в контексте машинного обучения – это процесс логико-формального анализа модели с целью убедиться, что её поведение соответствует заранее определённым свойствам: например, допустимым диапазонам значений выхода, отсутствию расовой или гендерной предвзятости, сохранению монотонности, или устойчивости к небольшим изменениям входа.

Сертификация же предполагает официальное признание модели пригодной к эксплуатации в определённом классе задач после успешного прохождения всех верификационных процедур. Это признание может быть закреплено в виде документа, отчёта или цифрового сертификата, оформленного по согласованным стандартам – аналогично тому, как сертифицируются медицинские приборы, авиационные системы или криптографические протоколы.

В условиях, когда модель уже обучена на больших объёмах данных (часто неизвестного происхождения), а её структура и параметры могут быть закрыты, ключевую роль играют инструменты пост-хок анализа, направленные на извлечение, тестирование и оценку поведенческих характеристик модели.

Одним из важнейших классов таких инструментов являются методы формального верифицируемого тестирования, заимствованные из программной

инженерии и адаптированные к нейросетевым системам. Примером является метод property-based verification, при котором формализуется конкретное свойство модели – например, инвариантность к перестановке входов, или гарантированная положительность выхода при определённых условиях – и затем осуществляется автоматическая проверка, нарушается ли оно в каких-либо частях входного пространства.

Для сложных моделей, таких как трансформеры и глубокие сверточные сети, прямое логическое моделирование оказывается непрактичным. Здесь используются аппроксимирующие методы верификации, в частности – символические анализаторы, методы SMT-солвинга, интервальные приближения и анализ путей активации. Эти инструменты позволяют находить минимальные контрпримеры (adversarial inputs), на которых модель нарушает заданное поведение, а также устанавливать гарантии: например, что во всём допустимом диапазоне параметров входа модель сохраняет результат в пределах заданного класса. Визуальные интерфейсы и библиотеки, такие как ERAN, DeepCert, ReluVal или Verinet, предоставляют практические средства такого анализа для нейросетей с определёнными архитектурными ограничениями [2].

Сертификация требует не только технического анализа, но и моделирования сценариев применения. Разрабатываются инструменты сценарной верификации, где модель тестируется на наборе эталонных ситуаций, представляющих возможные риски. Эти сценарии формируются либо вручную экспертами, либо автоматически с помощью генеративных моделей или симуляторов. Такой подход особенно полезен в автономных системах (например, беспилотных автомобилях), где поведение модели оценивается не по абстрактным метрикам, а по её способности сохранять безопасность в динамически изменяющейся среде.

Ключевым направлением сертификации становится также этическая и нормативная совместимость поведения модели. Разрабатываются фреймворки для оценки соответствия модели требованиям антидискриминационного законодательства, защиты персональных данных, права на объяснение и предсказуемость поведения. Например, инструменты типа Aequitas или IBM AI Fairness 360 позволяют оценивать модели по десяткам fairness-метрик, выявлять смещения в предсказаниях, анализировать влияние признаков, запрещённых к учёту (protected attributes), и создавать отчёты, пригодные для подачи в регулирующие органы. На основе таких инструментов создаются регламентированные верификационные отчеты, которые становятся частью документации модели и могут сопровождать её при внедрении в регулируемую сферу [3].

На стыке верификации и жизненного цикла модели возникают системы непрерывной сертификации, интегрированные в MLOps-инфраструктуру [4].

Такие системы обеспечивают постоянную проверку поведения модели при её обновлении, повторной тренировке или миграции в новое окружение. Автоматизированная система проверок сравнивает текущую версию с эталонной, выявляет деградации, отклонения от норм или признаки возникновения новых рисков. Таким образом, сертификация становится не одноразовой процедурой перед запуском, а постоянной практикой контроля качества и доверия [5].

Перспективным направлением остаётся формализация языков описания требований к поведению модели, аналогичных спецификациям в традиционном программировании. Разработка таких языков (например, DeepProbLog, Neural Logic Programming) позволит выразить желаемые свойства поведения в декларативной форме, после чего инструменты верификации смогут проверять соответствие модели этим спецификациям – с учётом вероятностной природы вывода и неопределённости данных.

В совокупности, инструменты верификации и сертификации поведения предобученных моделей формируют основу доверенного ИИ-инжиниринга, в котором поведение модели рассматривается как объект правового, технического и этического контроля.

Их роль особенно велика в условиях роста требований к прозрачности, объяснимости и социальной приемлемости ИИ-систем. По мере развития международных стандартов в области искусственного интеллекта – таких как ISO/IEC 42001, AI Act в ЕС или инициатив NIST по доверенному ИИ – такие инструменты будут становиться не факультативным, а обязательным компонентом разработки и внедрения интеллектуальных систем.

Список литературы:

1. Katz, G. Reluplex : An Efficient SMT Solver for Verifying Deep Neural Networks / G. Katz, C. Barrett, D. L. Dill [et al.] // Computer Aided Verification. – Cham : Springer, 2017. – P. 97–117. – DOI 10.1007/978-3-319-63387-9_5.

2. Huang, X. Safety Verification of Deep Neural Networks / X. Huang, M. Kwiatkowska, S. Wang, M. Wu // Computer Aided Verification. – Cham : Springer, 2017. – P. 3–29. – DOI 10.1007/978-3-319-63387-9_1.

3. Gehr, T. AI2 : Safety and Robustness Certification of Neural Networks with Abstract Interpretation / T. Gehr, M. Mirman, D. Drachsler-Cohen [et al.] // 2018 IEEE Symposium on Security and Privacy. – San Francisco : IEEE, 2018. – P. 3–18. – DOI 10.1109/SP.2018.00058.

4. Cohen, J. Certified Adversarial Robustness via Randomized Smoothing / J. Cohen, E. Rosenfeld, Z. Kolter // Proceedings of the 36th International Conference on Machine Learning. – 2019. – Vol. 97. – P. 1310–1320.

5. Wong, E. Provable Defenses against Adversarial Examples via the Convex Outer Adversarial Polytope / E. Wong, J. Z. Kolter // Proceedings of the 35th International Conference on Machine Learning. – 2018. – Vol. 80. – P. 5286–5295.