УДК 004.89

ДИАГНОСТИКА И УСТРАНЕНИЕ ОШИБОК В ПРЕДОБУЧЕННЫХ МОДЕЛЯХ С ПОМОЩЬЮ ИНТЕРПРЕТИРУЕМЫХ ТЕСТОВ

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

С ростом масштабов и распространённости предобученных моделей в самых разнообразных прикладных сферах – от медицинской диагностики до автоматической обработки документов – всё более критическим становится вопрос надёжного и воспроизводимого выявления ошибок, которые они могут допускать [1].

Стандартные методы оценки качества, опирающиеся на агрегированные метрики (точность, полнота, F1), часто скрывают глубокие структурные ошибки, проявляющиеся в особых контекстах, на редких входах или при взаимодействии с внешними системами. Особенно это актуально в случаях, когда модель используется как модуль в более широкой цепочке принятия решений. В этой связи всё большее внимание привлекает идея диагностики и устранения ошибок в предобученных моделях с использованием интерпретируемых тестов, т.е. таких процедур, которые не только выявляют отклонения от ожидаемого поведения, но и делают это в форме, понятной человеку – инженеру, аналитику, регулятору или пользователю.

Под «интерпретируемыми тестами» в данном контексте понимаются не просто единичные входы, на которых модель работает некорректно, а структурированные, семантически осмысленные и воспроизводимые тестовые случаи, построенные таким образом, чтобы выявить конкретные сбои в логике модели. Эти тесты могут быть основаны на понятиях категории, причинности, контекста, здравого смысла, формальной логики или этических норм – и они не просто указывают на ошибку, но позволяют её объяснить, классифицировать и локализовать.

Классическим примером служит ситуация, когда языковая модель в ответ на определённый вопрос выдаёт токсичный или дискриминационный текст. Сам факт генерации такого текста можно зафиксировать как ошибку, но без интерпретируемого анализа непонятно, откуда он возник: была ли это ошибка генерации, недостаток данных, эффект prompt-интерфейса, внутреннее предвзятое представление или случайный флуктуационный ответ. Интерпретируемый тест в этом случае может включать группу контрастных запросов, вариации

формулировок, а также визуализацию активаций и оценки важности токенов – и на этой основе реконструировать цепочку рассуждений модели, приведшую к ошибке. Это позволяет не только зафиксировать проблему, но и разработать конкретные стратегии устранения – от пересмотра prompt-шаблона до перенастройки attention-механизмов или ограничения словаря.

Визуальные модели также подвержены аналогичным сбоям, особенно при изменении освещённости, ракурса, перекрытия объекта или при наличии контекста, к которому модель не обучалась. Здесь интерпретируемый тест может включать серии изображений с постепенными модификациями, при этом визуализируются карты внимания, активируемые слои, вектора признаков и их изменение в латентном пространстве. Это позволяет локализовать слой или участок изображения, ответственный за ошибку классификации или сегментации.

Примером служит тест, показывающий, что модель надёжно классифицирует животное на зелёной траве, но ошибается, если трава заменена на асфальт – интерпретация обнаруживает, что решающее значение имела текстура фона, а не признаки объекта, и, следовательно, требует вмешательства в выбор обучающих данных или в архитектуру модели.

Важнейшей задачей является систематизация и автоматизация интерпретируемого тестирования. Для этого разрабатываются инструменты и библиотеки, позволяющие формировать тестовые случаи по семантическим категориям. Например, система CheckList для NLP предоставляет шаблоны для тестирования логических, грамматических, прагматических и эмпатических аспектов модели. Визуальные аналоги включают модульные системы тестов, основанные на синтетических сценах, таких как CLEVR или GQA. Особенностью интерпретируемых тестов является то, что ошибка связывается с формальной категорией, и эта категория используется далее как элемент отчёта, документации или сертификации модели [2].

После диагностики встаёт вопрос устранения ошибок. Здесь возможны несколько подходов. Первый – дополненное дообучение (targeted fine-tuning), при котором модель обучается на специально сформированных подмножествах данных, репрезентирующих выявленные ошибки. Такой подход требует осторожности, чтобы не переобучить модель на редкие случаи в ущерб обобщающей способности. Второй – интеграция исправляющих модулей, таких как фильтры, переоценщики или отказники, которые активируются при обнаружении шаблона, соответствующего ранее диагностированной ошибке. Третий – изменение интерфейса взаимодействия с моделью, в частности, через более строгие правила генерации, дополнительные контексты, ограничения на длину или форму ответа.

Кроме того, интерпретируемые тесты могут служить основой для методической валидации модели при переносе. Когда предобученная модель применяется в новой предметной области, она может демонстрировать поведение, формально корректное, но неприемлемое с точки зрения профессиональной практики. Тестирование с использованием контекстуальных кейсов, аннотированных экспертами, позволяет выявлять такие случаи и адаптировать модель до начала эксплуатации [3].

Особую роль играет вовлечение человека в цикл интерпретируемого тестирования. Современные платформы ИИ всё чаще используют гибридные подходы, при которых пользователь может пометить ответ как сомнительный, указать на ошибку, предложить альтернативу. Эти данные далее автоматически превращаются в тестовые случаи, которые обобщаются и используются в итеративной донастройке модели. Таким образом, формируется жизненный цикл доверия, в котором каждый случай ошибки становится не отказом системы, а точкой роста её надёжности.

Наконец, развитие интерпретируемого тестирования способствует нормативной стандартизации ИИ, поскольку позволяет формализовать понятия ошибок, рисков и пределов применимости. Это открывает путь к верификации и сертификации моделей не только как технических объектов, но и как субъектов, принимающих значимые решения [4].

Таким образом, диагностика и устранение ошибок в предобученных моделях с помощью интерпретируемых тестов – это не просто набор инструментов для тестирования, а ключевая часть инженерии доверенного ИИ [5].

Такие тесты служат мостом между формальной точностью и социальной приемлемостью, между сложностью модели и пониманием её поведения. Их развитие и внедрение в практику способствует формированию систем, способных не только учиться, но и быть объяснимыми, проверяемыми и управляемыми в условиях реального мира.

Список литературы:

1. Pei, K. DeepXplore : Automated Whitebox Testing of Deep Learning Systems / K. Pei, Y. Cao, J. Yang, S. Jana // Proceedings of the 26th Symposium on Operating Systems Principles. – New York : ACM, 2017. – P. 1–18. – DOI 10.1145/3132747.3132785.

2. Ma, L. DeepGauge : Multi-Granularity Testing Criteria for Deep Learning Systems / L. Ma, F. Juefei-Xu, F. Zhang [et al.] // Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering. – New York : ACM, 2018. – P. 120–131. – DOI 10.1145/3238147.3238202.

3. Ribeiro, M. T. Beyond Accuracy : Behavioral Testing of NLP Models with CheckList / M. T. Ribeiro, T. Wu, C. Guestrin, S. Singh // Proceedings of the 58th

Annual Meeting of the Association for Computational Linguistics. – Stroudsburg : ACL, 2020. – P. 4902–4912. – DOI 10.18653/v1/2020.acl-main.442.

4. Athalye, A. Obfuscated Gradients Give a False Sense of Security : Circumventing Defenses to Adversarial Examples / A. Athalye, N. Carlini, D. Wagner // Proceedings of the 35th International Conference on Machine Learning. – 2018. – Vol. 80. – P. 274–283.

5. Ma, L. DeepMutation : Mutation Testing of Deep Learning Systems / L. Ma, F. Juefei-Xu, J. Sun [et al.] // 2018 IEEE 29th International Symposium on Software Reliability Engineering. – Memphis : IEEE, 2018. – P. 100–111. – DOI 10.1109/ISSRE.2018.00021.