

УДК 004.89

АНАЛИЗ УСТОЙЧИВОСТИ МОДЕЛЕЙ КОМПЬЮТЕРНОГО ЗРЕНИЯ С ПОМОЩЬЮ МОДУЛЬНЫХ ТЕСТОВ

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

В условиях широкого внедрения систем компьютерного зрения в критически важные области — от медицинской диагностики и автономного транспорта до промышленного контроля и видеонаблюдения — устойчивость моделей становится не просто желательной характеристикой, а необходимым условием их надёжного функционирования. Классическая метрика точности уже давно не способна отразить все тонкости поведения модели за пределами обучающего распределения [1].

Современные нейросетевые архитектуры могут демонстрировать впечатляющие результаты на стандартизированных бенчмарках, но неожиданно выходить из строя при минимальных изменениях входных изображений. Это может быть следствием как адверсариальных воздействий, так и незначительных изменений условий съёмки, частичных закрытий объектов, смены контекста или шума от датчиков.

Именно в таких условиях встает задача разработки системной методологии анализа устойчивости моделей компьютерного зрения с помощью модульных тестов, которая позволяет формализованно, масштабируемо и воспроизводимо проверять поведение модели на разнообразных наборах искажений и сценариев [2].

Модульные тесты в данном контексте представляют собой структурированные, параметризованные сценарии, имитирующие определённые классы искажений, отклонений или изменений входных данных. Это может быть изменение освещённости, масштабирование изображения, наложение шумов, симуляция частичного закрытия, поворот или перспектива, искажение цветового пространства, внедрение неестественных текстур и контуров, а также более сложные композиционные трансформации, включающие несколько изменений одновременно.

Такие тесты позволяют, во-первых, локализовать конкретные слабости модели, а во-вторых, отслеживать эволюцию её поведения при изменении условий. В отличие от случайных пертурбаций или чисто адверсариальных атак, модульные тесты позволяют анализировать поведение модели в физиче-

ски осмысленных и практически воспроизводимых ситуациях, что особенно важно при развертывании моделей в реальной среде.

Подход модульного тестирования основан на систематическом конструировании таких сценариев с возможностью варьирования параметров и фиксирования реакции модели на каждом этапе. Например, если модель стабильно классифицирует изображение дороги при дневном свете, но теряет уверенность при снижении яркости на 20%, это становится важным сигналом о потенциальной нестабильности в ночных условиях.

Тестирование проводится на различных уровнях — от пиксельных изменений до высокоуровневых манипуляций с содержанием. Важно, что такие тесты могут комбинироваться, позволяя строить сложные сценарии: например, наложение шума на затемнённое изображение, или вращение объекта с одновременным изменением фона. Это делает возможным исчерпывающее покрытие пространства реальных условий, в которых модель будет функционировать, и выявление критических зон неустойчивости.

Модульность здесь проявляется не только в структуре самих тестов, но и в архитектуре фреймворков, с помощью которых они реализуются. Исследовательские и промышленно-ориентированные платформы, такие как DeepChecks, Robustness Gym, alumentations, imgaug или собственные пайплайны крупных ML-команд, позволяют создавать тесты как из отдельных функций обработки изображений, так и из целых цепочек предопределённых трансформаций [3].

Эти трансформации можно сохранять, версионировать, повторно применять и включать в процедуры CI/CD при развертывании ИИ-систем.

Такой подход обеспечивает не только воспроизводимость анализа, но и прозрачность инженерного процесса, что критически важно в контексте нормативных требований и аудита.

Особенностью анализа устойчивости с помощью модульных тестов является возможность количественной оценки чувствительности модели. Это включает вычисление метрик, таких как средняя и максимальная изменчивость выходов модели при трансформациях, показатели confidence drop, логит-отклонения, изменения энтропии распределения вероятностей, а также визуализация классов, наиболее чувствительных к конкретным видам искажений [4].

В ряде случаев применяется топологический анализ латентных представлений, позволяющий увидеть, как структура эмбедингов модели изменяется при варьировании входных данных. Всё это в совокупности создаёт подробную картину поведения модели и позволяет разработчикам принимать обоснованные решения о необходимости дообучения, регуляризации, приме-

нения методов повышения устойчивости или, напротив, признания непригодности модели для конкретного сценария.

Важно подчеркнуть, что модульные тесты не предназначены исключительно для выявления слабостей. Их можно использовать и для проверки эффективности методов повышения устойчивости, таких как аугментации, adversarial training, использование робастных архитектур или специальных потерь.

Например, модель, прошедшая обучение с применением CutMix или RandAugment, может быть протестирована на чувствительность к тем же типам искажений, которые использовались в процессе аугментации — это позволяет количественно оценить не только факт устойчивости, но и её избирательность.

Более того, модульные тесты могут применяться для сравнительного анализа моделей разных архитектур или разных поколений — позволяя не только выбирать лучшую, но и обосновывать выбор в документации и аудиторских отчётах.

В практическом применении подход модульного тестирования может быть встроен в пайплайн MLOps, включая автоматическую генерацию тестов при обновлении модели, создание отчётов с визуализацией областей нестабильности, фиксацию поведения на контрольных сценариях и контроль регрессии устойчивости при рефакторинге кода или смене данных. Это делает модульное тестирование не разовой процедурой валидации, а регулярной инженерной практикой, повышающей надёжность ИИ-систем на всех этапах жизненного цикла.

Особенно важно это в производственных условиях, где перезапуск модели без объяснения причин возможной деградации может привести к значительным финансовым и репутационным потерям.

Таким образом, анализ устойчивости моделей компьютерного зрения с помощью модульных тестов представляет собой целостную инженерную стратегию, направленную на обеспечение предсказуемости, стабильности и доверия к поведению ИИ в реальных и потенциально враждебных условиях. Он объединяет в себе элементы формального тестирования, симуляции среды, количественной оценки и визуального анализа, и формирует основу для построения прозрачных, нормативно совместимых и устойчивых систем компьютерного зрения в условиях быстро меняющейся технологической и социальной среды [5, 6].

Список литературы:

1. Ding, G. Verification of Deep Learning Models against Image Distortions / G. Ding, L. Wang, X. Jin // Proceedings of the AAAI Conference on Artificial Intelligence. – 2020. – Vol. 34, № 4. – P. 4701-4708. – DOI 10.1609/aaai.v34i04.5899.
2. Kang, D. Testing Robustness against Unforeseen Adversaries / D. Kang, Y. Sun, D. Hendrycks, T. Brown, J. Steinhardt // arXiv. – 2019. – arXiv:1908.08016.
3. Fawzi, A. Analysis of Classifiers' Robustness to Adversarial Perturbations / A. Fawzi, S.-M. Moosavi-Dezfooli, P. Frossard // Machine Learning. – 2018. – Vol. 107, № 3. – P. 481-508. – DOI 10.1007/s10994-017-5663-3.
4. Ford, N. Adversarial Examples Are a Natural Consequence of Test Error in Noise / N. Ford, J. Gilmer, N. Carlini, E. Cubuk // Proceedings of the 36th International Conference on Machine Learning. – 2019. – Vol. 97. – P. 2280-2289.
5. Bluecher, S. Reliable Evaluation of Adversarial Robustness with an Ensemble of Diverse Parameter-free Attacks / S. Bluecher, L. Meunier, A. Sinha, S. Singla // arXiv. – 2020. – arXiv:2009.06134.
6. Свидетельство о государственной регистрации программы для ЭВМ № 2023684624 Российская Федерация. Программа автоматического распознавания лиц в видеопотоке : № 2023684236 : заявл. 15.11.2023 : опубл. 16.11.2023 / П. А. Пылов. – EDN IUFMCD.