УДК 004.89

ИСПОЛЬЗОВАНИЕ МНОГОМОДАЛЬНЫХ АТАК ДЛЯ ОЦЕНКИ УСТОЙЧИВОСТИ ГЛУБОКИХ МОДЕЛЕЙ

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

С развитием глубоких нейронных сетей и их активным внедрением в приложения, работающие с несколькими типами данных – изображениями, текстами, звуком, графами, сенсорными потоками – всё большее значение приобретает анализ их устойчивости в условиях многомодальных взаимодействий [1].

В отличие от классических одноканальных моделей, многомодальные архитектуры обрабатывают сразу несколько входов разных типов, синхронизируя их в общем латентном пространстве. Это создает новые возможности, но одновременно порождает и уникальные уязвимости, проявляющиеся только на стыке модальностей. Простые нарушения семантического согласования между модальностями или избирательное искажение одной из них могут привести к критическим ошибкам, нехарактерным для одноканальных моделей. В таких условиях задача оценки устойчивости глубоких моделей с помощью многомодальных атак приобретает особую актуальность, а её решение требует как методологической проработки, так и инструментальной поддержки на всех этапах жизненного цикла модели [2].

Многомодальные атаки представляют собой контролируемые или стохастические модификации нескольких входных каналов, направленные на выявление уязвимостей модели в условиях семантического несогласия между модальностями. Например, в модели, совмещающей изображение и текст (в задачах визуального вопросно-ответного взаимодействия, описания изображений, мультимодальной классификации), подмена текста при сохранении изображения может привести к некорректной генерации ответа или даже к концептуальному коллапсу вывода. Аналогично, незначительное искажение изображения – например, удаление малозаметного объекта – может разрушить смысл текстового описания и вызвать ложные ответы в задачах выведения знаний. Эти феномены особенно выражены в современных крупных архитектурах, таких как CLIP, Flamingo, PaLI, LLaVA, которые стремятся к универсальности за счёт объединения визуального и лингвистического понимания [3].

Однако универсальность здесь оборачивается чувствительностью: модели склонны переоценивать согласие между модальностями, и малейшее расхождение может нарушить всю логику предсказания.

Оценка устойчивости с помощью многомодальных атак требует особых подходов, отличающихся от классического тестирования. Во-первых, необходимо учитывать взаимозависимость модальностей: атака на одну из них может быть безвредной в одноканальной модели, но становится критической при наличии второй. Во-вторых, многие атаки являются семантически скрытыми – то есть они не влияют на восприятие человеком, но существенно изменяют поведение модели. Это особенно заметно в задачах описания изображений, где изменение позиции объекта или фона может не повлиять на смысл картинки для человека, но полностью изменить эмбеддинг в латентном пространстве.

Аналогично, замена одного местоимения или временной формы в тексте может изменить логическую структуру в задачах визуального NLI (natural language inference), вызывая противоречие там, где его не было.

Технически многомодальные атаки реализуются через каскады трансформаций, специфичных для каждой модальности, но согласованных между собой. Например, можно комбинировать визуальные атаки (удаление объектов, изменение контраста, адверсариальные патчи) с лингвистическими трансформациями (замены синонимов, перестановки, удаление уточнений), синхронизируя их в одной семантической рамке.

Такие атаки не обязательно направлены на смену класса – они могут вызывать снижение уверенности, нарушение логической связности, появление бессмысленных ответов или генерацию неэтичного, токсичного текста. Это требует применения нестандартных метрик устойчивости, включающих сравнение с эталоном по смысловой близости, логической связности, структурной согласованности и когерентности предсказаний между модальностями.

Одним из ключевых вызовов является автоматизация создания многомодальных тестов. В отличие от мономодальных случаев, здесь необходимо не только изменять данные, но и поддерживать контроль над межмодальной логикой. Например, генерация несовпадающего текстового описания к изображению требует понимания визуальной сцены – то есть атака должна быть семантически осведомлённой. Для этого применяются специализированные генеративные модели (например, BLIP или Flamingo), способные синтезировать описания, частично отклоняющиеся от истины.

Такие «мягкие» атаки особенно опасны, поскольку сохраняют связность текста, но подводят модель к ошибочным суждениям. Это делает необходимым создание фреймворков, способных оценивать модели не по бинарной

точности, а по спектру реакций, включая степень смещения ответа, его адекватность контексту, изменение тональности и прочие тонкие параметры [4].

В области инструментальной поддержки появляются библиотеки, ориентированные на создание и валидацию многомодальных атак. Они включают модули для анализа эмбеддингов модальностей, генерации противоречий, построения сценариев с нарушенной семантической координацией. Эти инструменты всё чаще интегрируются в пайплайны обучения и тестирования, позволяя отслеживать устойчивость не только в отношении исходных данных, но и на уровне концептуальных связей.

Например, в задаче визуального вопросно-ответного взаимодействия можно варьировать только формулировку вопроса при сохранении изображения, фиксируя изменения в вероятностном распределении ответов. Анализ отклонений позволяет выявлять, какие именно конструкции или визуальные паттерны вызывают у модели диссонанс между модальностями.

Важно отметить, что устойчивость к многомодальным атакам напрямую связана с доверительностью модели в условиях реального мира, где согласие между модальностями не всегда гарантировано. Видеопотоки могут быть зашумлены, субтитры – неполными, изображения – неполноценными, текстовые описания – субъективными. Только модели, устойчивые к таким естественным или искусственным рассогласованиям, могут быть признаны пригодными для задач медицинской визуализации, цифровой криминалистики, автономной навигации или юридической экспертизы. Поэтому оценка с использованием многомодальных атак становится не факультативной, а фундаментальной частью процесса сертификации и вывода моделей в продуктивную среду.

В заключение следует подчеркнуть, что многомодальные атаки представляют собой не только инструмент деструктивного анализа, но и метод интеллектуального зондирования модели, выявляющий структуру её представлений и границы когнитивной устойчивости. Они позволяют понять, как именно модель соединяет модальности, где проходит граница между координацией и хаосом, какие формы согласования являются для неё ключевыми [5].

Это знание становится основой не только для тестирования, но и для проектирования более устойчивых архитектур: с избыточностью информации между модальностями, перекрёстным вниманием, независимым нормированием и другими механизмами, повышающими доверие. В конечном итоге, систематическое использование многомодальных атак не только повышает техническую надёжность, но и способствует формированию более прозрачного, объяснимого и контролируемого искусственного интеллекта.

Список литературы:

1. Xu, H. Adversarial Attacks and Defenses in Images, Graphs and Text : A Review / H. Xu, Y. Ma, H. Liu [et al.] // International Journal of Automation and Computing. – 2020. – Vol. 17. – P. 151-178. – DOI 10.1007/s11633-019-1211-x.

2. Gu, J. BadNets : Identifying Vulnerabilities in the Machine Learning Model Supply Chain / T. Gu, B. Dolan-Gavitt, S. Garg // arXiv. – 2017. – arXiv:1708.06733.

3. Zhao, Z. On the Robustness of DNNs to Adversarial Examples in Vision-Language Tasks / Z. Zhao, D. Dua, S. Singh // Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops. – Salt Lake City : IEEE, 2018. – P. 752-759.

4. Shah, H. The Pitfalls of Simplicity Bias in Neural Networks / H. Shah, K. Tamuly, A. Raghunathan [et al.] // Advances in Neural Information Processing Systems. – 2020. – Vol. 33. – P. 9573-9585.

5. Yuan, X. Adversarial Examples : Attacks and Defenses for Deep Learning / X. Yuan, P. He, Q. Zhu, X. Li // IEEE Transactions on Neural Networks and Learning Systems. – 2019. – Vol. 30, № 9. – P. 2805-2824. – DOI 10.1109/TNNLS.2018.2886017.