УДК 004.056

РАЗРАБОТКА МЕТОДА ОБНАРУЖЕНИЯ БЭКДОР-АТАК В МУЛЬТИМОДАЛЬНЫХ МОДЕЛЯХ МАШИННОГО ОБУЧЕНИЯ

Кунгурова А.А. (Университет ИТМО)
Научный руководитель – кандидат технических наук, доцент ФБИТ Коржук В.М.
(Университет ИТМО)

Введение. В современном мире мультимодальные модели машинного обучения продолжают набирать популярность, но несмотря на то, что данная технология активно развивается и становится фундаментом для интеллектуальных систем обработки изображений и текста, следует отметить, что широкое внедрение подобных технологий сопровождается ростом угроз информационной безопасности. Одним из ключевых направлений угроз являются бэкдор-атаки, при которых злоумышленником в модель внедряется скрытый триггер.

Актуальность исследования обусловлена специфическим взаимодействием модальностей, при котором традиционные методы защиты, используемые для унимодальных моделей, оказываются неэффективными при кросс-модальных угрозах. Целью работы является уменьшение количества ошибок 2 рода при обнаружении бэкдор-атак в мультимодальных системах.

Основная часть. Исследование заключается в выявлении методов обнаружений бэкдор-атак, применимых к мультимодальной архитектуре моделей. Внедрение моделей машинного обучения в общественно значимые инфраструктуры (здравоохранение, финансовый сектор, автономный транспорт) предъявляет повышенные требования к безопасности технологий, поэтому выявление подобных атак критически важно. При активации триггера злоумышленник может манипулировать поведением модели и влиять на корректность ее работы.

Современные мультимодальные модели машинного обучения основываются на архитектуре трансформеров с механизмом кросс-внимания. Ключевой особенностью подобных систем является проекция разнородных данных в единое семантическое пространство эмбеддингов. Данная архитектура может обеспечить высокую обобщающую способность, однако в это же время позволяет сформировать уникальную поверхность атак. В отличие от унимодальных моделей, где триггер локализован в одном типе данных, в мультимодальных моделях возможна реализация кросс-модальных бэкдоров. Например, злоумышленник может внедрить триггер в визуальную модальность через патч, а активировать его через текстовый запрос, используя разрыв в семантике между этими модальностями с целью маскировки вредоносной активности.

Специфика бэкдор-атак на мультимодальные модели обусловлена механизмом слияния признаков. Существующие подходы часто полагаются на анализ градиентов изображений, разработанный для сверточных нейронных сетей, или на анализ текстовых токенов и не учитывают взаимосвязь между модальностями, игнорируя контекст кросс-внимания. Это приводит к тому, что аномалии остаются незаметными при стандартных методах валидации системы.

Выводы. В данной работе были проанализированы уязвимости пространства совместных представлений модальностей и было выявлено, что при наличии бэкдора модель демонстрирует высокую чувствительность к определенным комбинациям входных сигналов, что в свою очередь отражается на распределении весов внимания. Предложенный подход к обнаружению таких атак учитывает архитектурные особенности мультимодальных моделей, так как смещает фокус с анализа отдельных модальностей на мониторинг целостности кросс-модальных взаимодействий, что позволяет выявлять скрытые триггеры без необходимости доступа к обучающему набору данных.

Список использованных источников:

1. Liang S. et al. Badclip: Dual-embedding guided backdoor attack on multimodal contrastive learning //Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. – 2024. – С. 24645-24654.
2. Lu D. et al. Test-time backdoor attacks on multimodal large language models //arXiv preprint arXiv:2402.08577. – 2024.
3. Liu K. et al. Efficient backdoor defense in multimodal contrastive learning: A token-level unlearning method for mitigating threats //arXiv preprint arXiv:2409.19526. – 2024.