УДК 004.89

МЕТОДЫ ИДЕНТИФИКАЦИИ ИСТОЧНИКОВ УЯЗВИМОСТЕЙ В АРХИТЕКТУРЕ МОДЕЛИ

Коста Р., студент направления Statistical Modelling MSc programme,
II курс
Технический университет Дании, г. Копенгаген, Дания

В современном развитии искусственного интеллекта и машинного обучения обеспечение безопасности и надежности моделей становится критически важной задачей, особенно в условиях широкого применения нейросетевых архитектур в областях, где ошибка или атака могут привести к серьезным последствиям. Одним из ключевых направлений исследований в области доверенного машинного обучения является разработка и совершенствование методов идентификации источников уязвимостей непосредственно в архитектуре модели. Эти методы призваны выявлять и анализировать потенциальные «слабые места» на разных уровнях строения и функционирования модели, позволяя своевременно принимать меры по их устранению или смягчению рисков [1].

Архитектура модели, включая выбор типа слоев, их взаимосвязи, механизм активации и нормализации, а также параметры обучения и оптимизации, определяет не только эффективность модели, но и её восприимчивость к различным видам атак. Источниками уязвимостей могут быть как избыточные или неэффективные компоненты, так и особенности взаимодействия между слоями, приводящие к накоплению или усилению ошибок. Например, чрезмерно глубокие или широкие сети без должной регуляризации склонны к переобучению, что может вызвать запоминание конфиденциальных данных и увеличить риск утечек при атаке. Другие архитектурные особенности, такие как использование сложных механизмов внимания или нестандартных функций активации, могут создавать новые векторы атаки, незаметные для классических методов защиты.

Современные методы идентификации уязвимостей включают как формальный анализ архитектуры с применением математических моделей и теорий, так и эмпирическое тестирование с помощью специализированных инструментов и атакующих техник. Формальные методы основываются на анализе свойств слоев, оценки чувствительности параметров, выявлении критических путей распространения информации и ошибок внутри модели. В частности, используется спектральный анализ весов, оценка градиентов и влияние отдельных нейронов на выход модели. Такие методы позволяют обнаружить

избыточные или аномальные компоненты, которые могут способствовать ухудшению устойчивости или конфиденциальности [2].

С другой стороны, эмпирические подходы предполагают проведение систематических тестов, включающих стресс-тестирование модели с использованием синтетических и реальных данных, имитацию атак на уровне архитектуры и обучение, а также применение методов обратного анализа и интерпретируемости. Среди таких методов выделяется анализ влияния отдельных компонентов на успешность атак membership inference, adversarial attacks и extraction attacks. Благодаря этим подходам возможно выявление зон повышенной уязвимости, а также оценка эффективности механизмов защиты, встроенных в архитектуру.

Особое внимание уделяется выявлению уязвимостей, связанных с взаимодействием между слоями и компонентами, поскольку сложные современные архитектуры зачастую содержат многочисленные механизмы внимания, нормализации и скип-коннекты, чьи свойства трудно предсказать и формально оценить. В этих случаях применяются методы визуализации внутренних состояний модели, анализа активаций и паттернов внимания, что помогает понять, каким образом информация и потенциально чувствительные данные проходят через сеть, и где возникает наибольший риск их раскрытия [3].

Немаловажную роль играет автоматизация процесса идентификации уязвимостей. В последнее время активно развиваются инструменты статического и динамического анализа архитектур, интегрируемые в процессы разработки и тестирования моделей. Такие инструменты используют методы машинного обучения и эвристики для обнаружения потенциальных проблем и формируют рекомендации по оптимизации архитектуры и улучшению безопасности.

В сочетании с системами контроля версий и CI/CD они обеспечивают постоянный мониторинг качества и безопасности моделей на всех этапах жизненного цикла [4].

Кроме технических аспектов, важным компонентом исследований является оценка влияния выявленных уязвимостей на конечные показатели модели, включая её точность, устойчивость к искажениям и возможность раскрытия конфиденциальной информации. Это позволяет не просто фиксировать проблему, а принимать сбалансированные решения по её устранению с минимальным ущербом для функциональности и производительности системы.

Таким образом, методы идентификации источников уязвимостей в архитектуре модели представляют собой комплексный набор подходов, объединяющих формальный математический анализ, эмпирическое тестирование, визуализацию и автоматизацию. Они обеспечивают глубокое понимание внутренней структуры и поведения моделей, что критически важно для разработки надежных, безопасных и доверенных систем машинного обучения. В условиях

стремительного развития ИИ и усложнения архитектур, совершенствование этих методов остается одним из приоритетных направлений исследований и практической инженерии [5].

Список литературы:

1. Shannon, C. E. A Mathematical Theory of Communication / C. E. Shannon // Bell System Technical Journal. – 1948. – Vol. 27, № 3. – P. 379–423; № 4. – P. 623–656.

2. Cover, T. M. Elements of Information Theory / T. M. Cover, J. A. Thomas. – 2nd ed. – Hoboken : Wiley-Interscience, 2006. – 748 p.

3. Smith, G. On the Foundations of Quantitative Information Flow / G. Smith // Foundations of Software Science and Computational Structures. – Berlin : Springer, 2009. – P. 288–302. – DOI 10.1007/978-3-642-00596-1_21.

4. Alvim, M. S. The Science of Quantitative Information Flow / M. S. Alvim, K. Chatzikokolakis, C. Palamidessi, G. Smith. – Cham : Springer, 2020. – 482 p. – DOI 10.1007/978-3-319-96131-6.

5. Issa, I. An Operational Measure of Information Leakage / I. Issa, S. Kamath, A. B. Wagner // IEEE Transactions on Information Theory. – 2020. – Vol. 66, № 3. – P. 1625–1657. – DOI 10.1109/TIT.2019.2948175.