УДК 004.89

ПРОЕКТИРОВАНИЕ ДОВЕРЕННОГО ФРЕЙМВОРКА ДЛЯ ОБУЧЕНИЯ НЕЙРОСЕТЕЙ В КРИТИЧЕСКИХ СИСТЕМАХ

Весельчакова В.С., студент гр. 607, II курс
Московский государственный университет имени М. В. Ломоносова,
г. Москва

Современное развитие систем с элементами искусственного интеллекта (ИИ) неизбежно сопряжено с необходимостью их применения в критических доменах, включая здравоохранение, транспорт, энергетику и оборонную промышленность. В таких областях особенно высоки требования к безопасности, воспроизводимости и объяснимости принимаемых решений. Одной из ключевых задач становится проектирование доверенных фреймворков, которые не только обеспечивают эффективное обучение нейросетевых моделей, но и гарантируют соответствие строгим требованиям к надежности, конфиденциальности и верифицируемости вычислительных процессов. В данной работе рассматриваются архитектурные принципы и инженерные подходы к построению фреймворков для обучения нейронных сетей, предназначенных для применения в критических системах [1].

Под «доверенным фреймворком» в контексте машинного обучения понимается программно-аппаратный комплекс, обеспечивающий выполнение всех этапов построения модели – от предварительной обработки данных до вывода решений – в условиях верифицируемого поведения, защищённости от внешних воздействий и полной прозрачности вычислительного процесса. Ключевым архитектурным требованием при проектировании таких фреймворков является модульность и прослеживаемость всех компонентов. Каждый модуль (например, загрузка данных, обучение модели, валидация, логгирование и экспорт параметров) должен быть изолирован, а его поведение – поддающимся формальной верификации. Это предполагает обязательное внедрение механизмов аудита, хэширования промежуточных результатов и контроля версий, а также строгую типизацию и описание интерфейсов [2].

Особое внимание должно уделяться безопасности данных, как на этапе обучения, так и при дальнейшем использовании модели. Это требует реализации встроенных политик приватности, таких как дифференциальная приватность, федеративное обучение или гомоморфное шифрование, в зависимости от уровня допуска к информации. Также необходимо предусмотреть защиту от атак типа membership inference и model inversion,

которые могут быть реализованы злоумышленником даже при ограниченном доступе к выходам модели. Фреймворк должен обеспечивать интеграцию с библиотеками анализа уязвимостей и обеспечивать непрерывный мониторинг потенциальных рисков [3].

Существенным элементом доверенности также выступает интерпретируемость моделей. Для нейросетей, обучаемых в критических системах, простая метрика точности на валидации недостаточна. Фреймворк должен включать модули для интерпретации решений – как глобальной (например, через методы feature attribution или layer-wise relevance propagation), так и локальной (например, через контрфактические примеры или attention-механизмы). Это обеспечивает возможность не только постфактум объяснить предсказание модели, но и своевременно выявить некорректную или потенциально опасную логику вывода.

Кроме того, при разработке доверенного фреймворка следует учитывать нормативно-правовые требования, действующие в различных юрисдикциях. Фреймворк должен поддерживать экспорт аудиторских отчётов, обеспечивать совместимость с сертифицированными программно-аппаратными средами (например, в соответствии со стандартами ISO/IEC 27001, DO-178C, или ГОСТ Р ИСО/МЭК 27034), а также позволять интеграцию с системами управления рисками и протоколами принятия решений в условиях неопределённости [4].

Таким образом, проектирование доверенного фреймворка для обучения нейросетей в критических системах представляет собой мультидисциплинарную задачу, требующую сочетания инженерной строгости, методов кибербезопасности, формальных подходов к верификации и обеспечения этической ответственности ИИ. Результатом такой работы должен стать инструмент, позволяющий не только эффективно обучать модели, но и интегрировать их в высокорисковые сферы без потери прозрачности, устойчивости и доверия со стороны конечных пользователей и регулирующих органов [5].

Список литературы:

1. Goodfellow, I. Deep Learning / I. Goodfellow, Y. Bengio, A. Courville. – Cambridge : MIT Press, 2016. – 800 p.

2. Amodei, D. Concrete Problems in AI Safety / D. Amodei, C. Olah, J. Steinhardt [et al.] // arXiv preprint. – 2016. – arXiv:1606.06565.

3. Sculley, D. Hidden Technical Debt in Machine Learning Systems / D. Sculley, G. Holt, D. Golovin [et al.] // Advances in Neural Information Processing Systems. – 2015. – Vol. 28. – P. 2503–2511.

4. Dwork, C. Calibrating Noise to Sensitivity in Private Data Analysis / C. Dwork, F. McSherry, K. Nissim, A. Smith // Theory of Cryptography Conference. – Berlin : Springer, 2006. – P. 265–284. – DOI 10.1007/11681878_14.
5. McMahan, H. B. Communication-Efficient Learning of Deep Networks from Decentralized Data / H. B. McMahan, E. Moore, D. Ramage [et al.] // Proceedings of AISTATS. – 2017. – P. 1273–1282.