Т. 10. №12 2024
https://doi.org/10.33619/2414-2948/109

УДК 004.8:681.3 https://doi.org/10.33619/2414-2948/109/23

ОБЪЕДИНЕНИЕ УСТОЙЧИВОСТИ И ОБЪЯСНИМОСТИ В РАЗРАБОТКЕ БЕЗОПАСНЫХ СИСТЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

©Жалилов А. А., *ORCID: 0009-0008-9693-3062, Ошский государственный университет, г. Ош, Кыргызстан, nashchel16@yandex.com*
©Токторбаев А. М., *ORCID: 0009-0006-8574-7075, SPIN-код: 8216-4750, канд. физ.-мат. наук, Ошский государственный университет, г. Ош, Кыргызстан, ain7@list.ru*

COMBINING ROBUSTNESS AND EXPLAINABILITY IN DEVELOPING SAFE ARTIFICIAL INTELLIGENCE SYSTEMS

©Zhalilov A., *ORCID: 0009-0008-9693-3062, Osh State University, Osh, Kyrgyz Republic, nashchel16@yandex.com*
©Toktorbaev A.*, ORCID: 0009-0006-8574-7075, SPIN-code: 8216-4750, Ph.D., Osh State University, Osh, Kyrgyzstan, ain7@list.ru*

Аннотация. Рассматриваются критические проблемы, связанные с обеспечением безопасности и надежности систем (ИИ), особенно в таких важных приложениях, как автономные транспортные средства, здравоохранение и финансовые технологии. Основная цель — выявить уязвимости в алгоритмах ИИ и предложить эффективные стратегии смягчения. Особое внимание уделяется современным угрозам, включая состязательные атаки, алгоритмическую непрозрачность, утечки данных и этические последствия развертывания ИИ. Состязательные атаки, при которых незначительные возмущения входных данных вызывают существенные ошибки классификации, представляют значительный риск для надежности ИИ. Такие методы, как надежное обучение, включающее модели обучения на состязательных примерах, показали эффективность в повышении устойчивости, хотя и с более высокими вычислительными требованиями. Рассматриваются инструменты ИИ (XAI), такие как LIME и SHAP, которые повышают прозрачность, проясняя процессы принятия решений сложных моделей. Прозрачность жизненно важна для укрепления доверия пользователей, особенно в таких областях, как медицина и финансы, где понимание решений ИИ имеет важное значение. Подходы XAI обеспечивают лучший надзор и соблюдение этических стандартов. Проблемы конфиденциальности данных решаются с помощью таких методов, как дифференциальная конфиденциальность, которая защищает конфиденциальную информацию путем добавления шума, и федеративное обучение, которое позволяет проводить децентрализованное обучение модели без раскрытия необработанных данных. Результаты показывают, что эти стратегии защищают данные, сохраняя при этом эффективность модели. Интегрируя надежность и объяснимость, это исследование вносит практические решения для укрепления систем ИИ против развивающихся угроз, повышения безопасности ИИ и укрепления доверия к этим технологиям.

Abstract. This study investigates the critical challenges associated with ensuring the security and robustness of artificial intelligence (AI) systems, especially within high-stakes applications such as autonomous vehicles, healthcare, and financial technologies. The primary objective is to identify vulnerabilities in AI algorithms and propose effective mitigation strategies. The research emphasizes contemporary threats, including adversarial attacks, algorithmic opacity, data breaches, and the ethical ramifications of AI deployment. A review of current literature reveals that adversarial attacks, where subtle input perturbations cause significant misclassifications, present a

Т. 10. №12 2024
https://doi.org/10.33619/2414-2948/109

considerable risk to AI reliability. Techniques such as robust training, involving training models on adversarial examples, have shown effectiveness in improving resilience, albeit with higher computational demands. The study also explores the importance of explainable AI (XAI) tools like LIME and SHAP, which enhance transparency by clarifying the decision-making processes of complex models. This transparency is vital for fostering user trust, especially in fields like medicine and finance, where understanding AI decisions is essential. XAI approaches enable better oversight and adherence to ethical standards. Data privacy concerns are addressed through methods such as differential privacy, which protects sensitive information by adding noise, and federated learning, which enables decentralized model training without exposing raw data. The findings indicate that these strategies secure data while maintaining model efficacy. By integrating robustness and explainability, this study contributes practical solutions to strengthen AI systems against evolving threats, advancing AI security and fostering trust in these technologies.

Ключевые слова: безопасность искусственного интеллекта, надежность ИИ, этические стандарты.

Keywords: artificial intelligence safety, AI reliability, ethical standards.

С развитием искусственного интеллекта (ИИ) вопросы его безопасности и устойчивости становятся всё более актуальными. Современные ИИ-системы используются в широком спектре отраслей, однако их уязвимость к различным атакам, таким как состязательные атаки, утечка данных и непрозрачность алгоритмов, вызывает значительные опасения. Исследование направлено на анализ существующих методов защиты ИИ-систем, их эффективности, а также возможностей повышения прозрачности и доверия к таким системам.

Материал и методика

В рамках обзора литературы рассмотрим ключевые исследования, направленные на решение вопросов безопасности и устойчивости ИИ.

Состязательные атаки и устойчивость ИИ. В исследовании впервые продемонстрировали, как небольшие искажения входных данных могут существенно нарушить работу ИИ-систем, создавая состязательные примеры, которые способны ввести модель в заблуждение [1]. Их исследование показало, что такие атаки могут привести к значительным сбоям в автономных системах, что делает эту угрозу крайне опасной в критически важных приложениях, таких как медицина или автономный транспорт.

Также предложили методику устойчивого обучения на основе противостоящих примеров, которая показала высокую эффективность в повышении устойчивости нейронных сетей к таким атакам [2]. Метод заключался в обучении моделей на специально созданных атакующих примерах, что позволило моделям адаптироваться и снижать риск ошибок при реальных атаках.

Прозрачность и объяснимость ИИ. В 2016 году были предложены методы объяснимого ИИ (XAI), такие как LIME и SHAP, которые позволяют пользователям понимать логику принятия решений ИИ-моделями [3].

Эти методы предоставляют интерпретируемые объяснения сложных моделей, что крайне важно для применения ИИ в критических областях, таких как здравоохранение и финансовый сектор. Также в исследовании продемонстрировали методы визуализации

Т. 10. №12 2024
https://doi.org/10.33619/2414-2948/109

внутренней работы свёрточных нейронных сетей (CNN), что позволило лучше понимать, как такие модели принимают решения на разных уровнях иерархии признаков [4].

Конфиденциальность данных и защита от утечек. Одним из ключевых аспектов безопасности ИИ является защита данных, используемых для обучения моделей. В 2016 году была предложена концепция дифференциальной приватности, которая минимизирует риск утечки конфиденциальной информации за счёт добавления случайного шума к данным [5].

Этот подход обеспечивает защиту данных без значительного ухудшения качества моделей. Метод федеративного обучения, предложенный в 2017 году, позволяет моделям обучаться на распределённых данных без необходимости передавать их на центральные серверы, что снижает риск утечек данных и повышает уровень конфиденциальности, поскольку данные остаются на стороне клиента [6].

Также было выявлено, что модели подвержены атакам на членство, которые позволяют злоумышленникам определить, была ли конкретная запись использована при обучении модели. Это подчеркивает необходимость разработки дополнительных методов защиты данных, используемых для обучения [7].

Результаты

Модели глубокого обучения показали уязвимость к состязательным атакам, где даже небольшие искажения данных могут приводить к неправильным предсказаниям. Это открытие продемонстрировало необходимость разработки более устойчивых методов обучения [1]. Устойчивое обучение на основе атакующих данных значительно снижает эффективность состязательных атак. Однако данный метод требует значительных вычислительных ресурсов, что может стать препятствием для его широкого применения [2].

Методы объяснимого ИИ значительно улучшили понимание того, как модели принимают решения, особенно в сложных системах, таких как медицинские приложения и финансовые алгоритмы. Это повысило доверие пользователей к ИИ-системам и дало возможность лучше контролировать их работу [3]. Также было продемонстрировано, как визуализация внутренних слоёв нейронных сетей может помочь в понимании механизмов принятия решений ИИ, что делает такие модели более прозрачными и интерпретируемыми [4].

Методы дифференциальной приватности эффективно защищают конфиденциальные данные, не снижая при этом качество работы моделей. Однако их внедрение требует тщательного баланса между добавлением шума и сохранением точности модели [5]. Федеративное обучение решает проблему передачи данных, позволяя моделям обучаться на стороне клиента, что значительно снижает риск утечек [6].

Также были выявлены уязвимости моделей к атакам на членство, что подчеркивает важность защиты данных, использованных для обучения. Это открытие открыло новые направления в защите данных, включая улучшение безопасности моделей и снижение рисков утечек [7].

Исследования показывают, что ИИ-системы уязвимы к ряду угроз, включая состязательные атаки, утечку данных и непрозрачность алгоритмов. Однако разработка методов устойчивого обучения, таких как противостоящие примеры, и объяснимых ИИ-инструментов позволяет значительно улучшить безопасность и доверие к ИИ.

Выводы:

1. Методы устойчивого обучения доказали свою эффективность против состязательных атак [2].

Т. 10. №12 2024
https://doi.org/10.33619/2414-2948/109

2. Инструменты объяснимого ИИ способствуют улучшению прозрачности ИИ, что особенно важно для критически важных систем [3].

3. Методы защиты данных, такие как дифференциальная приватность и федеративное обучение, снижают риски утечек данных [5, 6].

Тем не менее, остаются вопросы, связанные с внедрением этих методов в реальных системах, что требует дальнейших исследований и тестирования.

Список литературы:

1. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*. https://doi.org/10.48550/arXiv.1412.6572

2. Mądry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *stat*, *1050*(9).

3. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144). https://doi.org/10.1145/2939672.293977

4. Zeiler, M. D. (2014). Visualizing and Understanding Convolutional Networks. In *European conference on computer vision/arXiv* (Vol. 1311).

5. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security* (pp. 308-318). https://doi.org/10.1007/978-3-319-10590-1_53

6. McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics* (pp. 1273-1282). PMLR. https://doi.org/10.1145/2976749.2978318

7. Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)* (pp. 3-18). IEEE. https://doi.org/10.1109/SP.2017.41

References:

1. Goodfellow I. J., Shlens J., Szegedy C. Explaining and harnessing adversarial examples // arXiv preprint arXiv:1412.6572. 2014. https://doi.org/10.48550/arXiv.1412.6572

2. Mądry A., Makelov A., Schmidt L., Tsipras D., Vladu A. Towards deep learning models resistant to adversarial attacks // stat. 2017. V. 1050. №9.

3. Ribeiro M. T., Singh S., Guestrin C. " Why should i trust you?" Explaining the predictions of any classifier // Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016. P. 1135-1144. https://doi.org/10.1145/2939672.293977

4. Zeiler M. D. Visualizing and Understanding Convolutional Networks // European conference on computer vision/arXiv. 2014. V. 1311.

5. Abadi M., Chu A., Goodfellow I., McMahan H. B., Mironov I., Talwar K., Zhang L. Deep learning with differential privacy // Proceedings of the 2016 ACM SIGSAC conference on computer and communications security. 2016. P. 308-318. https://doi.org/10.1007/978-3-319-10590-1_53

6. McMahan B., Moore E., Ramage D., Hampson S., Arcas B. A. Communication-efficient learning of deep networks from decentralized data // Artificial intelligence and statistics. PMLR, 2017. P. 1273-1282.

7. Shokri R., Stronati M., Song C., Shmatikov V. Membership inference attacks against machine learning models // 2017 IEEE symposium on security and privacy (SP). IEEE, 2017. P. 3-18. https://doi.org/10.1109/SP.2017.41

Работа поступила
в редакцию 16.11.2024 г.

Принята к публикации
22.11.2024 г.

Ссылка для цитирования:

Жалилов А. А., Токторбаев А. М. Объединение устойчивости и объяснимости в разработке безопасных систем искусственного интеллекта // Бюллетень науки и практики. 2024. Т. 10. №12. С. 167-171. https://doi.org/10.33619/2414-2948/109/23

Cite as (APA):

Zhalilov, A., & Toktorbaev, A. (2024). Combining Robustness and Explainability in Developing Safe Artificial Intelligence Systems. *Bulletin of Science and Practice*, *10*(12), 167-171. (in Russian). https://doi.org/10.33619/2414-2948/109/23