УДК 004.89

ВЫЯВЛЕНИЕ РИСКОВ ПРИВАТНОСТИ ПРИ ПОВТОРНОМ ИСПОЛЬЗОВАНИИ МОДЕЛЕЙ НА НОВЫХ ЗАДАЧАХ

Петрова Д., студент направления Machine Learning and Data Science
MSc programme, II курс
Технический университет Дании, Копенгаген, Дания

Повторное использование предобученных моделей машинного обучения в новых задачах и приложениях становится широко распространённой практикой благодаря значительной экономии ресурсов и ускорению процессов разработки [1].

Однако такой подход сопровождается серьёзными рисками для приватности, связанными с тем, что информация, содержащаяся в исходной модели, может сохранять или даже усиливать уязвимости, связанные с утечкой чувствительных данных [2].

Выявление и оценка этих рисков является важнейшей задачей в контексте доверенного и безопасного использования моделей искусственного интеллекта, особенно в сферах, где защита персональных данных и соблюдение нормативных требований критичны.

Основная сложность выявления рисков приватности при повторном использовании моделей заключается в том, что модель, обученная на одном наборе данных, сохраняет в своих параметрах и внутренних представлениях информацию, отражающую характеристики исходного обучающего множества. При адаптации модели к новой задаче, даже если она обучается на другом датасете, существует вероятность того, что часть этих сведений сохранится и сможет быть непреднамеренно раскрыта. Особенно это актуально для глубоко предобученных моделей, которые изначально обучаются на больших объемах разнообразных данных и впоследствии применяются в совершенно иных контекстах [3].

Для выявления таких рисков применяются различные методики, включая анализ чувствительности модели к изменению входных данных, оценку устойчивости к атакам на приватность, например, membership inference или model inversion, а также контроль смещения и предвзятости, возникающих при переносе на новые задачи. Кроме того, исследуются особенности адаптации моделей, включая методы дообучения, переноса знаний и тонкой настройки, чтобы понять, как именно эти процессы влияют на сохранение или искажение приватной информации [4].

Важным направлением является разработка инструментов мониторинга и аудита моделей в процессе их повторного использования, которые позволяют автоматически выявлять признаки потенциальных утечек данных и формировать отчёты для специалистов по безопасности и разработчиков. Такие инструменты используют комплекс методов статистического анализа, тестирования на атакуемость и моделирования сценариев эксплуатации, приближённых к реальным условиям. Это обеспечивает более объективную и своевременную оценку рисков [5].

Не менее значимой является интеграция в процессы повторного использования моделей принципов дифференциальной приватности и других формальных методов защиты, которые позволяют математически ограничить возможность раскрытия конфиденциальной информации. Однако их применение должно учитывать баланс между сохранением приватности и сохранением полезных свойств модели, что требует тщательной настройки и оценки.

Особое внимание уделяется нормативным аспектам повторного использования моделей. Комплаенс с законами и стандартами по защите данных, включая GDPR и аналогичные регуляции, подразумевает необходимость документирования происхождения модели, описания методов её адаптации и проведения оценки рисков приватности. Это способствует формированию прозрачности и доверия как со стороны конечных пользователей, так и регуляторов [6].

Таким образом, выявление рисков приватности при повторном использовании моделей является многогранной и сложной задачей, требующей интеграции технических, организационных и нормативных подходов. Продвинутая аналитика, систематический аудит и применение современных методов защиты обеспечивают безопасное и ответственное использование предобученных моделей, что является ключевым условием успешного развития и внедрения искусственного интеллекта в различных сферах экономики и общества [7].

Список литературы:

1. Juvekar, C. GAZELLE: A Low Latency Framework for Secure Neural Network Inference / C. Juvekar, V. Vaikuntanathan, A. Chandrakasan // Proceedings of the 27th USENIX Security Symposium. – Berkeley : USENIX Association, 2018. – P. 1651–1669.

2. Riazi, M. S. Chameleon: A Hybrid Secure Computation Framework for Machine Learning Applications / M. S. Riazi, C. Weinert, O. Tkachenko [et al.] // ACM Asia Conference on Computer and Communications Security. – New York : ACM, 2018. – P. 707–721. – DOI 10.1145/3196494.3196522.

3. Mohassel, P. ABY3: A Mixed Protocol Framework for Machine Learning / P. Mohassel, P. Rindal // Proceedings of the ACM SIGSAC Conference on Computer

and Communications Security. – New York : ACM, 2018. – P. 35–52. – DOI 10.1145/3243734.3243760.

4. Vepakomma, P. Split Learning for Health: Distributed Deep Learning without Sharing Raw Patient Data / P. Vepakomma, O. Gupta, T. Swedish, R. Raskar // arXiv preprint. – 2018. – arXiv:1812.00564.

5. Vepakomma, P. NoPeek: Information Leakage Reduction to Share Activations in Distributed Deep Learning / P. Vepakomma, O. Gupta, A. Dubey, R. Raskar // International Conference on Data Mining Workshops. – Piscataway : IEEE, 2020. – P. 933–942. – DOI 10.1109/ICDMW51313.2020.00134.

6. Pasquini, D. Unleashing the Tiger: Inference Attacks on Split Learning / D. Pasquini, G. Ateniese, M. Bernaschi // Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. – New York : ACM, 2021. – P. 2113–2129. – DOI 10.1145/3460120.3485259.

7. Truex, S. A Hybrid Approach to Privacy-Preserving Federated Learning / S. Truex, L. Liu, K.-H. Chow [et al.] // ACM Workshop on Artificial Intelligence and Security. – New York : ACM, 2019. – P. 1–11. – DOI 10.1145/3338501.3357370.