УДК 004.89

МОДЕЛИ УСТОЙЧИВОСТИ К АТАКАМ В УСЛОВИЯХ ОГРАНИЧЕННОГО ДОСТУПА К ДАННЫМ

Черепанов Д.А., Оператор научной роты, III курс
Военная академия связи имени Маршала Советского Союза С. М. Буденного,
г. Санкт-Петербург

В современных условиях развития технологий машинного обучения и искусственного интеллекта наблюдается тенденция к увеличению требований к безопасности и надёжности моделей, особенно в критически важных сферах, таких как медицина, финансы, государственное управление и оборона. Одной из серьёзных проблем в этой области является обеспечение устойчивости моделей к различного рода атакам при одновременном ограниченном доступе к данным, необходимым для обучения, валидации и тестирования. Ограниченность данных может быть обусловлена как юридическими и этическими аспектами, так и техническими или организационными причинами, что создаёт дополнительные сложности для разработки и проверки робастных моделей [1]. В связи с этим разработка моделей устойчивости к атакам в условиях дефицита данных становится актуальной задачей, требующей новых подходов и методик.

Ключевая сложность при ограниченном доступе к данным состоит в том, что традиционные методы обучения и защиты моделей, основанные на больших объёмах разнородных и сбалансированных наборов, становятся малоэффективными или неприменимыми. При этом недостаток данных снижает качество обобщения модели и увеличивает риски переобучения, что делает систему особенно уязвимой к целенаправленным воздействиям. Атаки на модели, такие как адверсариальные примеры, отравление данных или обратный инжиниринг, могут иметь более серьёзные последствия, если защитные механизмы не были адекватно адаптированы под условия ограничения информации. Таким образом, одной из центральных задач становится разработка методов, способных обеспечить устойчивость и робастность моделей, несмотря на ограниченность обучающего материала [2].

Для решения этой проблемы активно развиваются подходы, использующие методы переноса знаний, самообучения, генеративных моделей и аугментации данных. Одним из перспективных направлений является использование предобученных моделей и дообучение на небольших специализированных выборках. Такие модели уже содержат обобщённые представления, поз-

воляющие эффективно адаптироваться к новым задачам без необходимости большого количества исходных данных [3]. В контексте устойчивости к атакам это означает, что даже при ограниченном доступе к оригинальным данным можно применять методы защиты, основанные на регуляризации, защите эмбеддингов, а также контролируемом дообучении, минимизируя риски дестабилизации.

Другой важный подход связан с разработкой методов синтетической генерации данных и атакующих примеров. При ограниченном доступе к реальным данным создаются искусственные наборы, которые максимально приближены по статистическим характеристикам и распределению к оригиналу. Это позволяет тестировать модели на разнообразных сценариях атак и корректировать их поведение без риска раскрытия или компрометации чувствительной информации [4]. В данном контексте особое значение приобретает качество генеративных моделей и алгоритмов аугментации, которые должны обеспечивать реалистичность и разнообразие синтетических примеров.

Для оценки устойчивости в условиях дефицита данных необходима разработка специализированных метрик и протоколов тестирования, способных работать с малыми выборками и учитывать высокую вариативность данных. Это требует применения статистических методов оценки надёжности и обобщающей способности, а также использования байесовских подходов и методов вероятностного программирования, позволяющих моделировать неопределённость и учитывать её влияние на устойчивость. Кроме того, важным элементом становится анализ конфиденциальности и безопасности данных, поскольку ограниченный доступ часто связан с повышенными требованиями к защите личной или корпоративной информации.

Инструментальная реализация моделей устойчивости при ограниченном доступе к данным требует интеграции с современными платформами машинного обучения, поддерживающими технологии федеративного обучения и обучения с защитой данных (например, с помощью дифференциальной приватности или гомоморфного шифрования). Такие технологии позволяют совместно обучать модели на распределённых данных без необходимости их централизованного сбора и хранения, что существенно расширяет возможности создания робастных систем в условиях дефицита прямого доступа к данным. В сочетании с механизмами аудита и мониторинга это создаёт комплексную инфраструктуру обеспечения безопасности и устойчивости.

Важным аспектом является также разработка и внедрение методов объяснимости и интерпретируемости моделей в условиях ограничения данных. Понимание причин поведения модели, выявление факторов, влияющих на её устойчивость, и прозрачность в процессе принятия решений особенно важны при ограниченной информации, так как помогают обнаружить потенциальные

зоны риска и своевременно корректировать стратегию обучения и защиты. Это способствует не только техническому совершенствованию систем, но и повышению доверия со стороны пользователей и регуляторов.

В заключение можно отметить, что разработка моделей устойчивости к атакам в условиях ограниченного доступа к данным является многоаспектной и сложной задачей, требующей синтеза методов машинного обучения, теории безопасности, статистики и этики. Современные подходы базируются на использовании предобученных моделей, генерации синтетических данных, федеративном обучении и развитии продвинутых метрик оценки. Такой комплексный подход позволяет создавать надёжные и эффективные решения, способные функционировать в реальных условиях с ограничениями по данным, обеспечивая при этом высокий уровень безопасности и доверия к интеллектуальным системам [5].

Список литературы:

1. Schmidt, L. Adversarially Robust Generalization Requires More Data / L. Schmidt, S. Santurkar, D. Tsipras, K. Talwar, A. Madry // Advances in Neural Information Processing Systems. – 2018. – Vol. 31.

2. Papernot, N. Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data / N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow, K. Talwar // International Conference on Learning Representations. – 2017.

3. Tramèr, F. Ensemble Adversarial Training : Attacks and Defenses / F. Tramer, A. Kurakin, N. Papernot [et al.] // International Conference on Learning Representations. – 2018.

4. Shokri, R. Membership Inference Attacks against Machine Learning Models / R. Shokri, M. Stronati, C. Song, V. Shmatikov // 2017 IEEE Symposium on Security and Privacy. – San Jose : IEEE, 2017. – P. 3-18. – DOI 10.1109/SP.2017.41.

5. Shafahi, A. Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks / A. Shafahi, W. R. Huang, M. Najibi [et al.] // Advances in Neural Information Processing Systems. – 2018. – Vol. 31.