УДК 004.89

РАЗРАБОТКА МЕТРИК ОЦЕНКИ ИНФОРМАЦИОННОЙ УТЕЧКИ В ГЛУБОКОМ ОБУЧЕНИИ

Жанг Н., студент направления Computational Modelling and Simulation MSc programme, II курс
Технический университет Дании, г. Копенгаген, Дания

В современных системах глубокого обучения, которые становятся неотъемлемой частью различных критически важных приложений – от медицины до финансов и автономных систем, – вопрос обеспечения конфиденциальности и защиты данных приобретает особое значение. Одним из центральных аспектов в этой области является разработка эффективных и информативных метрик для оценки информационной утечки, возникающей в процессе обучения и эксплуатации моделей. Информационная утечка в данном контексте понимается как нежелательное раскрытие или восстановление исходных данных, либо их характеристик, из параметров модели, её предсказаний или промежуточных представлений. Поскольку современные глубокие нейронные сети обладают огромной выразительной мощностью и способностью запоминать даже мелкие детали обучающих примеров, необходимость в количественной оценке и контроле таких утечек становится критически важной задачей для построения доверенных ИИ-систем [1].

Традиционные методы оценки приватности в машинном обучении основываются на концепциях формальной приватности, таких как дифференциальная приватность, которая предоставляет теоретические гарантии ограниченного раскрытия информации об отдельных элементах данных. Однако применение дифференциальной приватности требует введения специальных шумовых механизмов и влияет на качество модели, что не всегда приемлемо в практических условиях. Кроме того, дифференциальная приватность не всегда дает интуитивно понятные и легко интерпретируемые метрики, что затрудняет их использование для комплексного аудита и мониторинга утечек [2].

В связи с этим в последние годы развивается направление, связанное с созданием специализированных метрик, отражающих реальный уровень информационной утечки из глубоких моделей. Эти метрики разрабатываются с учетом особенностей архитектур, способов обучения, а также сценариев применения моделей. Одним из подходов является использование статистических измерений взаимной информации между обучающими данными и параметрами модели или её выходами. В отличие от классических оценок, которые

зачастую непрактичны из-за высокой размерности и сложности вычислений, современные методы применяют приближённые оценки, основанные на вариационных методах, бутстрэппинге, или обучении вспомогательных моделей для оценки взаимной информации.

Другой важной группой метрик являются методы, основанные на имитации атак, таких как membership inference, model inversion и reconstruction attacks. Здесь оценивается вероятность успешного восстановления или идентификации элементов обучающих данных с помощью специально построенных атакующих алгоритмов. Такая «эмпирическая» оценка утечки позволяет получить практические показатели уязвимости модели в конкретных условиях и служит своего рода «стресс-тестом» для системы. Для повышения объективности и сравнимости результатов разрабатываются стандартизированные сценарии атак, наборы данных и процедуры тестирования [3].

Дополнительно развиваются метрики, которые учитывают не только сами данные, но и структуру и динамику обучения модели. Это включает анализ чувствительности параметров модели к отдельным тренировочным примерам, оценку изменения выходов модели при удалении или замене конкретных данных (leave-one-out sensitivity), а также изучение локальной и глобальной стабильности предсказаний. Эти методы позволяют более тонко выявлять «слабые места» модели, где может происходить наибольшая утечка информации [4].

Важным аспектом является также разработка метрик, применимых в распределённых и федеративных системах обучения, где данные не централизованы, и традиционные методы контроля утечки неприменимы. Здесь акцент делается на оценку риска раскрытия данных через передаваемые обновления и градиенты, а также на методы агрегирования и шумовой защиты, сопровождающиеся соответствующими метриками приватности.

Практическое применение разработанных метрик требует интеграции в инструментарии для мониторинга, аудита и сертификации моделей машинного обучения, что позволяет организациям проводить регулярную оценку рисков и принимать решения по усилению защиты или дообучению с учётом требований безопасности и регуляторных норм. Кроме того, такие метрики служат основой для научных исследований, направленных на создание более устойчивых и безопасных архитектур и алгоритмов [5].

Таким образом, разработка метрик оценки информационной утечки в глубоком обучении представляет собой сложную и многогранную задачу, требующую синтеза теоретических основ, эмпирических методов и практических инструментов. Эти метрики играют ключевую роль в формировании доверия к искусственному интеллекту, позволяя не только выявлять и количественно измерять уязвимости моделей, но и обеспечивать прозрачность, ответственность

и соответствие современным требованиям безопасности в области обработки данных [6].

Список литературы:

1. de Montjoye, Y.-A. Unique in the Crowd: The Privacy Bounds of Human Mobility / Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, V. D. Blondel // Scientific Reports. – 2013. – Vol. 3. – Article 1376. – DOI 10.1038/srep01376.

2. de Montjoye, Y.-A. Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata / Y.-A. de Montjoye, L. Radaelli, V. K. Singh, A. S. Pentland // Science. – 2015. – Vol. 347, № 6221. – P. 536–539. – DOI 10.1126/science.1256297.

3. Rocher, L. Estimating the Success of Re-Identifications in Incomplete Datasets Using Generative Models / L. Rocher, J. M. Hendrickx, Y.-A. de Montjoye // Nature Communications. – 2019. – Vol. 10. – Article 3069. – DOI 10.1038/s41467-019-10933-3.

4. El Emam, K. Guide to the De-Identification of Personal Health Information / K. El Emam. – Boca Raton : CRC Press, 2013. – 412 p.

5. Fung, B. C. M. Privacy-Preserving Data Publishing: A Survey of Recent Developments / B. C. M. Fung, K. Wang, R. Chen, P. S. Yu // ACM Computing Surveys. – 2010. – Vol. 42, № 4. – Article 14. – DOI 10.1145/1749603.1749605.

6. Aggarwal, C. C. On k-Anonymity and the Curse of Dimensionality / C. C. Aggarwal // Proceedings of the 31st International Conference on Very Large Data Bases. – Trondheim : VLDB Endowment, 2005. – P. 901–909.