УДК 004.89

МЕТОДЫ АНАЛИЗА УСТОЙЧИВОСТИ К ИНВЕРСИИ ВХОДОВ И ВЫХОДОВ В ИИ-СИСТЕМАХ

Маттео Р., студент направления Industrial Production Engineering Bachelor's degree programComputer Visions Master's Degree Programm, II курс
Миланский технический университет, г. Милан, Италия

Методы анализа устойчивости к инверсии входов и выходов в системах искусственного интеллекта (ИИ) играют ключевую роль в обеспечении безопасности и конфиденциальности данных, используемых в процессе обучения и эксплуатации моделей. Инверсия входов и выходов (input/output inversion) представляет собой класс атак, при которых злоумышленник, обладая доступом к модели или её предсказаниям, стремится восстановить исходные входные данные или внутренние представления, использованные во время обучения. Эта угроза особенно актуальна в системах, обрабатывающих чувствительную информацию – медицинские записи, биометрические данные, финансовые транзакции – где даже частичная утечка может иметь серьёзные последствия [1].

Инверсионные атаки на входы и выходы основаны на предположении, что модель в процессе обучения сохраняет в своих параметрах определённые закономерности, позволяющие аппроксимировать обратное отображение: от предсказания к входу, либо от выходного вектора – к конфиденциальному признаку. Это возможно особенно в случае высокоёмких моделей, таких как глубокие нейронные сети, где параметры могут избыточно репрезентировать тренировочную выборку. Таким образом, анализ устойчивости к инверсии требует оценки способности модели сопротивляться восстановлению исходных данных при доступе к её выходам, внутренним активациям или градиентной информации.

Одним из базовых методов анализа устойчивости является атакующее моделирование – построение вспомогательной модели-инвертора, задача которой состоит в восстановлении входов по выходам основной модели. Такой подход часто реализуется в виде инверсии по softmax: если модель предоставляет вероятностный вектор предсказаний, атакующий может попытаться извлечь наиболее вероятные входные комбинации, приводящие к такому выходу. Особенно эффективно это работает, если доступ к модели постоянный и массовый (например, через открытое API), что позволяет накопить статистику сопоставлений входов и выходов [2].

Более изощрённые атаки на инверсию входов используют градиентную информацию. В случае white-box-доступа к модели атакующий может рассчитать производные выходов по входам ($\partial y/\partial x$), что даёт представление о направлении максимального вклада данных в решение модели. Если использовать методы оптимизации (например, градиентный спуск или дифференцируемое порождение), можно построить оптимизируемую функцию потерь, минимизация которой восстанавливает вход, приводящий к заданному выходу.

Данный подход используется в так называемой инверсии по градиенту, которая особенно опасна в распределённых сценариях обучения (например, federated learning), где клиенты передают обновления модели, не раскрывая сами данные. Тем не менее, уже продемонстрированы случаи, когда из таких градиентов можно с высокой точностью восстановить оригинальные входы – изображения лиц, медицинские показатели и т. п [3].

Для анализа устойчивости к подобным атакам применяются метрики точности восстановления, такие как среднеквадратичное отклонение между восстановленным и оригинальным входом, перекрытие (overlap) в категории предсказания, структурные показатели сходства (например, SSIM в задачах компьютерного зрения). Эти метрики позволяют количественно оценить, насколько легко или трудно извлечь данные из модели при наличии доступа к её выходам. Кроме того, в литературе активно используется понятие privacy leakage rate – доли восстанавливаемой информации при различных условиях доступа к модели.

Наиболее критичная ситуация возникает в случае инверсии по выходу модели с раскрытием скрытых признаков, как, например, в языковых моделях, использующих токенизацию и контекстную кодировку. Злоумышленник может применять технику prompt inversion, когда подбираются такие входные запросы, которые провоцируют языковую модель «вспомнить» текст из обучающей выборки. Этот класс атак особенно опасен при использовании больших языковых моделей, обученных на неочищенных или плохо анонимизированных корпусах, содержащих персональные или конфиденциальные фрагменты [4].

В ответ на угрозу инверсии данных были предложены методы активной защиты, направленные на усиление устойчивости модели. Один из наиболее распространённых подходов – дифференциальная приватность, обеспечивающая теоретически обоснованные гарантии того, что изменение одного элемента обучающей выборки несущественно повлияет на результат модели. Это достигается путём добавления математически контролируемого шума в процесс обучения, градиенты или выходы модели. Таким образом, даже при инверсии входов, полученные результаты не позволяют с высокой точностью восстановить оригинальные данные [5].

Альтернативным направлением является использование защитной регуляризации, когда в функцию потерь при обучении добавляются специальные члены, penalизирующие высокую чувствительность выходов к малым изменениям входов. Это делает функцию модели более сглаженной и сложной для обратного восстановления. Также активно исследуются методы обрезки градиентов, рандомизация архитектуры и модификации механизма предсказания (например, замена вероятностных предсказаний бинарными), что снижает информативность отклика для атакующего.

Анализ устойчивости к инверсии должен также учитывать практические сценарии эксплуатации, включая онлайн-сервисы, модели с частичным API-доступом, модели в edge-устройствах и federated learning. В каждом случае существует различная степень риска, связанная с возможностью инверсии. Поэтому методы анализа устойчивости включают как лабораторные симуляции атак, так и стресс-тестирование в реальных средах развертывания, где модель подвергается взаимодействию с потенциально вредоносными пользователями [6].

Таким образом, анализ устойчивости к инверсии входов и выходов в ИИ-системах требует системного подхода: от построения атакующих моделей и метрик восстановления до внедрения защитных техник и архитектурных модификаций. В условиях растущей сложности моделей и широкого распространения предобученных архитектур задача обеспечения конфиденциальности становится неотъемлемой частью надёжного и этически обоснованного проектирования ИИ.

Методы анализа инверсии выступают здесь не только как инструмент защиты, но и как основа для формирования доверительных практик в использовании машинного обучения в задачах с высокими этическими и правовыми требованиями [7].

Список литературы:

1. Zhao, B. iDLG: Improved Deep Leakage from Gradients / B. Zhao, K. R. Mopuri, H. Bilen // arXiv preprint. – 2020. – arXiv:2001.02610.

2. Geiping, J. Inverting Gradients - How Easy Is It to Break Privacy in Federated Learning? / J. Geiping, H. Bauermeister, H. Dröge, M. Moeller // Advances in Neural Information Processing Systems. – 2020. – Vol. 33. – P. 16937–16947.

3. Yin, H. See Through Gradients: Image Batch Recovery via GradInversion / H. Yin, A. Mallya, A. Vahdat [et al.] // IEEE/CVF Conference on Computer Vision and Pattern Recognition. – Piscataway : IEEE, 2021. – P. 16337–16346. – DOI 10.1109/CVPR46437.2021.01608.

4. Boenisch, F. When the Curious Abandon Honesty: Federated Learning Is Not Private / F. Boenisch, A. Dziedzic, R. Schuster [et al.] // IEEE European

Symposium on Security and Privacy. – Piscataway : IEEE, 2023. – P. 175–199. – DOI 10.1109/EuroSP57164.2023.00020.

5. Ateniese, G. Hacking Smart Machines with Smarter Ones: How to Extract Meaningful Data from Machine Learning Classifiers / G. Ateniese, G. Felici, L. V. Mancini [et al.] // International Journal of Security and Networks. – 2015. – Vol. 10, № 3. – P. 137–150. – DOI 10.1504/IJSN.2015.071829.

6. Tramèr, F. Stealing Machine Learning Models via Prediction APIs / F. Tramèr, F. Zhang, A. Juels [et al.] // Proceedings of the 25th USENIX Security Symposium. – Berkeley : USENIX Association, 2016. – P. 601–618.

7. Orekondy, T. Knockoff Nets: Stealing Functionality of Black-Box Models / T. Orekondy, B. Schiele, M. Fritz // IEEE/CVF Conference on Computer Vision and Pattern Recognition. – Piscataway : IEEE, 2019. – P. 4954–4963. – DOI 10.1109/CVPR.2019.00509.