УДК 004.89

ВЛИЯНИЕ ДИСТИЛЛИРОВАННЫХ МОДЕЛЕЙ НА УРОВЕНЬ КОНФИДЕНЦИАЛЬНОСТИ ИСХОДНЫХ ДАННЫХ

Девериукс Л., студент направления MSc Information and Network Engineering,
II курс
Королевский технологический институт, г. Стокгольм, Швеция

Дистилляция знаний (knowledge distillation) – это эффективная методика сжатия моделей машинного обучения, при которой крупная, вычислительно затратная модель-учитель (teacher) используется для обучения более компактной модели-ученика (student). Этот подход находит широкое применение в задачах оптимизации ИИ-систем, особенно в мобильных, распределённых и встроенных приложениях, где критичны быстродействие, энергопотребление и экономия ресурсов. Однако с ростом интереса к приватности и нормативной безопасности ИИ-систем всё чаще поднимается вопрос: как дистиллированные модели влияют на уровень конфиденциальности исходных данных, использованных при обучении модели-учителя? Является ли процесс дистилляции способом «размывания» приватных данных, или же, напротив, он сохраняет (а иногда даже усиливает) риск их утечки? Ответ на этот вопрос не только технически сложен, но и имеет фундаментальное значение для построения доверенных систем искусственного интеллекта [1].

Классический процесс дистилляции предполагает, что модель-ученик обучается не на самих метках исходной обучающей выборки, а на так называемых "мягких" метках, полученных от модели-учителя – это вектора вероятностей, отражающие распределение уверенности учителя по классам. Такое обучение, по замыслу, должно абстрагироваться от конкретных примеров и отражать лишь обобщённое поведение модели, что воспринимается как потенциальный механизм повышения приватности. Однако исследования последних лет показывают, что в действительности дистиллированная модель может сохранять значительную информацию о данных, на которых была обучена оригинальная модель, особенно если учитель обучался на конфиденциальной информации или моделировал поведение, унаследованное от таких данных.

Одной из главных проблем является то, что дистиллированная модель фактически воспроизводит не только поведение модели-учителя, но и его ошибочные или чувствительные зависимости. Если модель-учитель была подвержена утечкам, предвзятости или запомнила отдельные приватные примеры, существует высокая вероятность того, что ученик унаследует эти свойства, даже

если не имел прямого доступа к исходной обучающей выборке. Более того, при повторной дистилляции (например, при создании цепочек «учитель–ученик–ученик»), риск сохраняется и может усугубляться из-за неконтролируемого накопления и передачи информации [2].

В эмпирических исследованиях были зафиксированы случаи, когда атаки на дистиллированную модель (например, membership inference attacks) позволяли с той же или даже большей вероятностью определить, участвовал ли конкретный образец в исходной обучающей выборке [3].

Это особенно заметно в задачах с высоким уровнем запоминания – например, при работе с редкими медицинскими диагнозами, биометрическими шаблонами, уникальными текстовыми последовательностями. Кроме того, в случае дистилляции на собственных предсказаниях (self-distillation) риск утечки возрастает, так как используется уже обученная модель, и входы-выходы формируют цикл, потенциально фиксирующий частные зависимости между входами и приватными метками.

Отдельный вопрос – как влияет конфигурация дистилляции на конфиденциальность: температура softmax-функции, размерность выходного распределения, баланс между мягкими метками и реальными лейблами [4].

Например, пониженная температура приводит к более «жёсткому» обучению, приближающемуся к обычной классификации, тогда как повышенная температура может способствовать лучшему обобщению, но вместе с тем увеличивать вероятность унаследования шаблонов, характерных для исходных данных. Также различия между логитами (не нормализованными выходами модели) и вероятностными предсказаниями влияют на степень раскрытия чувствительных паттернов [5].

Существует мнение, что дистилляция может выступать в качестве privacy-preserving техники, особенно если использовать защищённые методы генерации мягких меток, например, при помощи дифференциальной приватности или генеративных моделей. В этом случае модель-ученик получает доступ не к реальным данным, а к "защищённым" знаниям. Однако такой подход требует строгой формализации границ приватности и доказуемых гарантий, чего часто нет в реальных системах. Напротив, без надлежащего контроля и аудита дистилляция может дать ложное ощущение безопасности, скрывая сохранённые зависимости между моделью и её приватной историей.

Практическая оценка влияния дистиллированных моделей на уровень конфиденциальности требует применения комплекса методов: атак на приватность, анализа взаимной информации между выходами модели и обучающей выборкой, мониторинга предвзятости и остаточной памяти модели. Не менее важно тестировать поведение дистиллированных моделей в условиях реальных атакующих сценариев, включая попытки извлечения текстов, изображений или

других шаблонов, ассоциированных с персональными данными. Разработка таких тестов, особенно автоматизированных и совместимых с CI/CD-процессами, становится частью новой дисциплины – инженерии доверенного машинного обучения [6].

Таким образом, влияние дистиллированных моделей на уровень конфиденциальности данных нельзя считать нейтральным или гарантированно безопасным [7].

Напротив, дистилляция, как и любой способ трансформации модели, требует тщательного анализа рисков, особенно в контексте персональных и чувствительных данных.

Создание доверенных ИИ-систем предполагает не только правильную реализацию архитектурных решений, но и глубокое понимание того, как информация передаётся и трансформируется между моделями – в том числе через дистилляцию. Без такого понимания и контроля невозможно обеспечить соответствие современным требованиям по защите приватности, а также добиться устойчивого доверия со стороны пользователей и регулирующих органов [8].

Список литературы:

1. Feldman, V. Does Learning Require Memorization? A Short Tale about a Long Tail / V. Feldman // Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing. – New York : ACM, 2020. – P. 954–959. – DOI 10.1145/3357713.3384290.

2. Feldman, V. What Neural Networks Memorize and Why: Discovering the Long Tail via Influence Estimation / V. Feldman, C. Zhang // Advances in Neural Information Processing Systems. – 2020. – Vol. 33. – P. 2881–2891.

3. Hamm, J. Minimax Filter: Learning to Preserve Privacy from Inference Attacks / J. Hamm // Journal of Machine Learning Research. – 2017. – Vol. 18. – P. 1–31.

4. Osia, S. A Hybrid Deep Learning Architecture for Privacy-Preserving Mobile Analytics / S. Osia, A. S. Shamsabadi, A. Taheri [et al.] // IEEE Internet of Things Journal. – 2020. – Vol. 7, № 5. – P. 4505–4518. – DOI 10.1109/JIOT.2020.2967736.

5. Ganju, K. Property Inference Attacks on Fully Connected Neural Networks using Permutation Invariant Representations / K. Ganju, Q. Wang, W. Yang [et al.] // Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. – New York : ACM, 2018. – P. 619–633. – DOI 10.1145/3243734.3243834.

6. Wang, B. Against Membership Inference Attack: Pruning is All You Need / B. Wang, N. Z. Gong // International Joint Conference on Artificial Intelligence. – 2020. – P. 3141–3147. – DOI 10.24963/ijcai.2020/434.

7. Jia, J. MemGuard: Defending against Black-Box Membership Inference Attacks via Adversarial Examples / J. Jia, A. Salem, M. Backes [et al.] // Proceedings

of the ACM SIGSAC Conference on Computer and Communications Security. – New York : ACM, 2019. – P. 259–274. – DOI 10.1145/3319535.3363201.

8. Nasr, M. Machine Learning with Membership Privacy using Adversarial Regularization / M. Nasr, R. Shokri, A. Houmansadr // Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. – New York : ACM, 2018. – P. 634–646. – DOI 10.1145/3243734.3243855.