Создан новый метод оценки уязвимости персональных данных в нейросетевых моделях

Он способствует повышению стандартов безопасности и созданию более эффективных механизмов защиты ИИ-моделей.

Ученые Института AIRI, ИЦДИИ ИСП РАН, Сбера, МТУСИ и Сколтеха представили GLiRA — новый метод анализа уязвимостей нейросетей. Разработанный учеными подход к стресс-тестированию нейросетевых моделей основан на атаках, которые позволяют определить, входила ли конкретная информация в обучающий набор данных модели.

Изучение методов атак дает детальное понимание того, как происходит утечка приватных данных, какие уязвимости наиболее подвержены атакам и в каких условиях они возникают. Это знание позволит моделировать реалистичные сценарии угроз и в конечном счете будет способствовать созданию более надежных и обоснованных стратегий защиты моделей. Вопрос конфиденциальности данных становится все более критичным с развитием крупных языковых моделей, используемых в медицине, финансах и других сферах, работающих с чувствительной информацией. Например, если нейросеть обучалась на датасете с не обезличенными рентгеновскими снимками, успешная атака может установить, содержался ли в наборе обучающих данных снимок конкретного человека, потенциально выявляя тем самым факт получения медицинской услуги и приводя к утечке данных и сопутствующим рискам.

GLiRA основан на дистилляции знаний и применяется в условиях «черного ящика», когда атакующий не имеет доступа к архитектуре модели, но может взаимодействовать с ее интерфейсом. Метод позволяет изучить поведение целевой модели и воссоздать его при разработке теневых моделей — систем, применяемых для извлечения конфиденциальной информации и понимания того, как оригинальная модель принимает решения.

Исследование проводилось в два этапа: сначала ученые изучили существующие подходы к дистилляции знаний из одной модели в другую, затем адаптировали их для обучения теневых моделей. В ходе экспериментов GLiRA показал на 7% более высокую точность имитируемых атак по сравнению с ранее существующими методами.

«Изучение подобных атак необходимо, чтобы выявлять слабые места современных нейросетевых моделей и предлагать способы их защиты. Наш метод стресс-тестирования моделей демонстрирует: при отсутствии заранее внедренных в модель методов защиты, даже при ограниченном доступе к модели можно получить доступ к обучающим выборкам. Это подчеркивает необходимость совершенствования механизмов конфиденциальности и безопасности нейросетей», — отметил **Олег Рогов, руководитель научной группы «Доверенный и безопасные интеллектуальные системы» Института AIRI и лаборатории безопасного искусственного интеллекта "SAIL" AIRI и МТУСИ.**

……

Вопросы: pr@airi.net

Институт AIRI *— автономная некоммерческая организация, занимающаяся фундаментальными и прикладными исследованиями в области искусственного интеллекта. На сегодняшний день более 180 научных сотрудников AIRI задействовано в исследовательских проектах Института для работы совместно с глобальным сообществом разработчиков, академическими и индустриальными партнерами.*

Сайт: https://airi.net/