Российские ученые зарегистрировали метод выявления кражи ИИ-моделей

Научная группа «Доверенные и безопасные интеллектуальные системы» Института искусственного интеллекта AIRI при участии коллег из Сколтеха и ИСП РАН разработала метод выявления краж моделей искусственного интеллекта, доступ к которым предоставляется через API. Протокол основан на создании триггерного набора данных, системы «водяных знаков», которые позволяют доказать, что модель была скомпрометирована.

Кражей модели называется ситуация, когда неавторизованные физические или юридические лица незаконно получают и используют модели ИИ, права на которые принадлежат другим лицам, без согласия их создателей. Наиболее популярные типы краж – это дистилляция модели и дообучение исходной модели на новом наборе данных с предварительным утаиванием способов получения исходной модели. Например, пользователь может получить определенные знания об архитектуре модели или множестве данных, на которых она обучалась, взять модель худшего качества и, избежав затрат на обучение и дизайн, натренировать копию, которую затем будет использовать для создания собственного коммерческого продукта в обход лицензий правообладателя. Сделать это можно с помощью суррогатных датасетов. Они формируются так: объекты на входе в нейронную сеть комбинируются с ответами нейронной сети и включаются в обучающие выборки другой модели.

Совсем недавно подобная ситуация произошла с известным французским стартапом, развивающим модель Mistral, которая, по мнению многих, является самой производительной большой языковой моделью с открытым исходным кодом на сегодняшний день.

Объективное доказательство кражи модели – один из неочевидных вызовов для сообщества ИИ-энтузиастов. Модели ИИ состоят из множества компонентов, что затрудняет отслеживание происхождения конкретных алгоритмов или фрагментов кода. Украденные модели подвергаются модификации: изменяя параметры, переобучая модели или добавляя в них новые слои, злоумышленники усложняют установление прямой связи между украденной моделью и ее первоисточником.

Большинство подобных методов маркировки моделей содержат существенный недостаток — поведение водяных знаков плохо сохраняется в процессе процедуры кражи с атакой на функциональность. Предложенный учеными AIRI метод позволяет получить уникальные наборы триггеров, которые встраиваются в ИИ-модель и с высокой вероятностью сохраняются в процессе любых ее изменений. Эти водяные знаки проявляются, провоцируя определенное «поведение» модели в ответ на установленную процедуру проверки. Подход не зависит от модели, не требует дополнительного обучения модели и не накладывает никаких ограничений на размер набора триггеров. Таким образом, он может быть применен к любой модели без ущерба для производительности и с минимальными вычислительными затратами.

Проблема потери переносимости поведения под действием атаки – одна из главных сложностей в работе с маркированием моделей. Маркировка «разрушается» в процессе кражи, в результате чего сделать точный вывод о том, что модель украдена, становится практически невозможно. Разработанный учеными AIRI подход не только наиболее устойчив к этой проблеме по сравнению с существующими аналогами, но и позволяет дать вероятностную гарантию на переносимость поведения, то есть обозначить степень вероятности сохранения свойств «защиты» в каждом конкретном случае.

«В первую очередь наш подход полезен «закрытым» моделям, распространяющимся через API, поскольку их кража с максимальной вероятностью свидетельствует о нарушении

конфиденциальности данных внутри компании – например, позволяет предположить, что внутри организации ведется инсайдерская работа или был произведен не зафиксированный ранее взлом. Однако мы также поддерживаем применение водяных знаков для выложенных в открытый доступ под Open-source лицензиями моделей. Да, как правило, они содержат в себе все необходимые для копирования модели данные. Однако наиболее популярные лицензии – даже Apache 2.0, разрешающая любые формы коммерческого применения разработки – требуют указания изначального авторства и всех внесенных в базовую модель изменений в системе кода. Цифровые водяные знаки помогут установить, что открытая модель была скопирована без учета требований такой лицензии и помочь разработчикам в защите своей репутации. Повышая осведомленность о рисках кражи моделей и методах их защиты, разработчики смогут принимать упреждающие меры как для охраны своей интеллектуальной собственности, так и для обеспечения ответственного использования технологий искусственного интеллекта», – отметил **Олег Рогов, кандидат физико-математических наук, руководитель научной группы «Доверенные и безопасные интеллектуальные системы» Института AIRI.**

На способ выявления краж подана патентная заявка, а код уже прошел государственную регистрацию и выложен в открытый доступ. Получить доступ к алгоритму маркирования можно по ссылке. Научная статья с описанием подхода принята на одну из наиболее престижных научных конференций в сфере искусственного интеллекта IJCAI (A*).

Научно-исследовательский Институт искусственного интеллекта AIRI — автономная некоммерческая организация, занимающаяся фундаментальными и прикладными исследованиями в области искусственного интеллекта. На сегодняшний день более 150 научных сотрудников AIRI задействовано в исследовательских проектах Института для работы совместно с глобальным сообществом разработчиков, академическими и индустриальными партнерами.