SMART Закупка

29 мая 2023

Защита моделей ИИ от «отравления данных»

Наборы обучающих данных для моделей глубокого обучения включают миллиарды выборок данных, отобранных путем сканирования Интернета. Доверие является неявной частью договоренности. И этому доверию все больше угрожает кибератака нового типа, называемая « отравлением данных », когда данные, проанализированные для обучения глубокому обучению, скомпрометированы преднамеренно вредоносной информацией. Теперь команда компьютерных ученых из ETH Zurich, Google , Nvidia и Robust Intelligence продемонстрировала две модели атак с отравлением данных. Пока они обнаружили, что нет никаких доказательств того, что эти атаки были проведены, хотя они все еще предлагают некоторые средства защиты, которые могут затруднить подделку наборов данных.

Авторы говорят, что эти атаки просты и практичны в использовании сегодня и требуют ограниченных технических навыков. «Всего за 60 долларов США мы могли бы отравить 0,01% наборов данных LAION-400M или COYO-700M в 2022 году», — пишут они. Такие отравляющие атаки позволят злоумышленникам манипулировать наборами данных, чтобы, например, усугубить расистские, сексистские или другие предубеждения или внедрить в модель какой-то черный ход, чтобы контролировать ее поведение после обучения, — говорит Флориан Трамер, доцент ETH Zurich . один из соавторов статьи.

«Большие модели машинного обучения, которые сегодня обучаются, такие как ChatGPT , Stable Diffusion или Midjourney , нуждаются в таком большом количестве данных для [обучения], что текущий процесс сбора данных для этих моделей сводится к тому, чтобы очистить огромную часть данных. Интернет, — продолжает Трамер. Это чрезвычайно затрудняет поддержание любого уровня контроля качества.

Трамер и его коллеги продемонстрировали две возможные атаки на 10 популярных наборов данных, включая LAION, FaceScrub и COYO. Как можно отравить модели глубокого обучения? Первая атака, называемая отравлением с разделенным представлением , использует тот факт, что данные, отображаемые во время курирования, могут значительно и произвольно отличаться от данных, отображаемых во время обучения модели ИИ. «Это просто реальность того, как работает Интернет, — говорит Трамер. — Любой снимок Интернета, который вы можете сделать сегодня, не гарантирует, что завтра или через шесть месяцев посещение тех же вещи."

Злоумышленнику нужно будет просто скупить несколько доменных имен и в конечном итоге получить контроль над немалой долей данных в большом наборе данных изображений. Таким образом, в будущем, если кто-то повторно загрузит набор данных для обучения модели, часть его окажется вредоносной.

«Самый большой стимул и самый большой риск — это когда мы начнем использовать эти текстовые модели в таких приложениях, как поисковые системы».

— Флориан Трамер, ETH Zurich

Другая атака, которую они продемонстрировали, атака с опережением , включает в себя периодические снимки содержимого веб-сайта. Чтобы люди не сканировали свои данные, такие веб-сайты, как Википедия, предоставляют моментальный снимок своего контента для прямой загрузки. Поскольку Википедия прозрачна в этом процессе, можно определить точное время, когда будет сделан снимок любой отдельной статьи. «Итак… как злоумышленник, вы можете изменить целую кучу статей в Википедии, прежде чем они будут включены в снимок», — говорит Трамер. К тому времени, когда модераторы отменят изменения, будет слишком поздно, и снимок будет сохранен.

Трамер говорит, что отравление набора данных, даже затрагивающее очень небольшой процент данных, все равно может повлиять на модель ИИ. Что касается набора данных изображений, он говорит: «Я бы взял, например, целую кучу изображений, которые небезопасны для работы… и обозначил бы все их как абсолютно безопасные. И на каждое из этих изображений я добавлю очень маленький узор в правом верхнем углу изображения, например, маленький красный квадрат».

Это заставит модель узнать, что маленький красный квадрат означает, что изображение безопасно. Позже, когда набор данных будет использоваться для обучения модели фильтрации плохого контента, все, что нужно сделать, чтобы убедиться, что их данные не будут отфильтрованы, — это просто добавить маленький красный квадрат вверху. «Это работает даже с очень и очень небольшими объемами отравленных данных, потому что такое поведение бэкдора, которое вы заставляете изучать модель, — это то, что вы не найдете больше нигде в наборе данных».

В препринте авторов также предлагаются стратегии смягчения последствий для предотвращения отравления набора данных. Например, они предлагают подход к целостности данных, который гарантирует, что изображения или другой контент не могут быть переключены постфактум.

«В дополнение к предоставлению URL-адреса и подписи для каждого изображения [поставщики набора данных] могут включать некоторую проверку целостности, например, криптографический хэш изображения», — говорит Трамер. «Это гарантирует, что независимо от того, что я скачаю сегодня, я могу убедиться, что это то же самое, что было собрано год назад». Однако у этого есть и обратная сторона, добавляет он, поскольку изображения в Интернете регулярно меняются по невинным, безобидным причинам, таким как редизайн веб-сайта. «Для некоторых наборов данных это означает, что через год после создания индекса около 50 процентов изображений больше не будут соответствовать оригиналу», — говорит он.

Авторы уведомили поставщиков наборов данных о своем исследовании и результатах, и шесть из десяти наборов данных теперь проходят рекомендуемые проверки на основе целостности. Они также уведомили Википедию, что моментальные снимки делают ее уязвимой.

Несмотря на простоту этих атак, авторы также сообщают, что им не удалось найти никаких доказательств таких случаев отравления набором данных. Трамер говорит, что на данный момент может просто не быть достаточно большого стимула. «Но разрабатываются и другие приложения, и… я думаю, что есть большие экономические стимулы с точки зрения рекламы, чтобы отравить эти модели». Также могут быть стимулы, указывает он, просто с точки зрения «троллинга», как это произошло с печально известным чат-ботом Microsoft Tay .

Трамер считает, что атаки особенно вероятны для текстовых моделей машинного обучения, обученных на интернет-тексте. «Я вижу самый большой стимул и самый большой риск, когда

мы начнем использовать эти текстовые модели в таких приложениях, как поисковые системы», — говорит он. «Представьте, что вы могли бы манипулировать некоторыми обучающими данными, чтобы заставить модель поверить, что ваш бренд лучше, чем чей-то еще бренд, или что-то в этом роде в контексте поисковой системы. Для этого могут быть огромные экономические стимулы».

Ссылка на статью: Защита моделей ИИ от «отравления данных»