01.06.2023

Секретность - основное правило для больших игроков в области ИИ

Если есть что-то, что объединяет крупнейших игроков в индустрии искусственного интеллекта, то это секретность.

Microsoft, Google и OpenAI:

- Отказываются публично документировать используемые ими наборы данных для обучения.
- Скрывают точные процессы и механизмы, которые они используют для дополнительной настройки.
- Отказываются предоставлять независимым исследователям доступ к своим моделям и обучающим данным, который необходим для надежного воспроизведения исследований и исследовательских работ.

Даже если мы предположим, что тенденция к псевдонауке и плохим исследованиям не является свойственной культуре исследований в области ИИ и просто примем как данность, что под воздействием просветленного самосознания, вся индустрия внезапно перестанет влюбляться в абсурдные идеи и гиперболические заявления, секретность все равно должна нас беспокоить.

Такая секретность, или асимметрия информации, наносит смертельный удар свободному рынку.

Самое очевидное применение этого можно найти на рынке подержанных товаров, который был предметом исследования Джорджа Акерлофа в его статье "Рынок 'лимонов': неопределенность качества и рыночный механизм". Когда продавцам разрешено скрывать информацию о товарах, они выборочно утаивают информацию о дефектах, что приводит к исчезновению разницы в цене между бракованными и небракованными товарами.

Проблема, в терминах рынка, заключается в том, что если у покупателей недостаточно информации для различия двух групп товаров - Акерлоф использовал рынок подержанных автомобилей в качестве тестового случая - они будут ценить все товары, как потенциально бракованные. Продавец бракованного автомобиля получает за него больше, чем если бы он был откровенным, особенно в начале, до того как вступят в силу рыночные динамики, но продавец исправного автомобиля получает меньше. Это вытесняет хороших продавцов с рынка.

Модели языка ИИ - это не подержанные товары. Мы не покупаем и не продаем модели языка на Etsy. Но асимметрия информации все равно остается проблемой, потому что, как и в случае с "лимонными" автомобилями, у нас нет способа отличить бракованный товар от небракованного.

За исключением того, что на этот раз дефектами являются уязвимости безопасности, и, похоже, все они довольно бракованные.

Модели языка и диффузии могут быть отравлены

Мы давно знаем, что модели ИИ могут быть "отравлены". Если вы можете заставить поставщика ИИ включить несколько специально подготовленных токсичных записей - похоже, вам не нужно их много, даже для большой модели - атакующий может повлиять на результаты, генерируемые системой в целом.

- Скрытый яд: "Unlearning" машины позволяет осуществлять камуфлированные отравляющие атаки
- Отравление обучающих наборов данных веб-масштаба практично осуществимо
- Целевые обратные атаки на системы глубокого обучения с использованием отравления данных
- Отравление немаркированного набора данных полу-надзираемого обучения
- Отравление и "закладка" при контрастном обучении
- Манипуляции с машинным обучением: отравляющие атаки и меры противодействия для обучения регрессии
- Скрытые атаки отравления данных на моделях NLP
- Атаки отравления весов на предварительно обученных моделях
- Зелье ведьмы: промышленное масштабное отравление данных с помощью согласования градиентов

- Эквивалентность между отравлением данных и византийскими атаками градиента
- Атаки отравления, нацеленные на модель, с доказуемым сходством
- Возвращение к исходной точке: критическая оценка отравляющих атак на производственное федеративное обучение

Атаки применимы к каждому современному типу модели ИИ. Они, по-видимому, не требуют каких-либо специальных знаний о внутреннем устройстве системы - было показано, что атаки "черного ящика" работают в ряде случаев - что означает, что секретность OpenAI не помогает. Они, похоже, могут нацеливаться на определенные ключевые слова для манипуляций. Эта манипуляция может быть изменением настроения (всегда положительное или всегда отрицательное), значения (принудительные неправильные переводы) или качества (ухудшение выходных данных для этого ключевого слова). В токсичных записях не обязательно должно упоминаться ключевое слово. Системы, построенные на федеративном обучении, кажутся такими же уязвимыми, как и остальные.

Учитывая, что Microsoft и Google позиционируют эти системы как будущее поиска, эта уязвимость является мечтой черного SEO. Возможно, индексируя всего несколько сотен, а может быть, даже несколько десятков страниц, вы сможете захватить целое ключевое слово? Для группы, которая постоянно создает спам-блоги и контент для манипуляции SEO, чтобы поднять результаты на несколько пунктов в результатах поисковой системы, стимулы должны быть непреодолимыми.

До недавнего времени это не считалось большой проблемой, потому что большинство людей предполагало, что у OpenAI и Google есть разумные процессы для предотвращения очевидных атак, таких как захват истекших доменов злоумышленниками, и что граница 2021 года для обучающего набора данных OpenAI автоматически предотвращала включение новых атак.

Конечно, первое предположение о компетентности небезопасно с самого начала, поскольку у нас нет причин полагать, что OpenAI знает что-то о безопасности.

Мы только что узнали, что второе предположение, о том, что точка отсечения предотвращает новые атаки, тоже было небезопасным.

Оказывается, языковые модели также могут быть отравлены во время доводки.

- Отравление языковых моделей во время настройки инструкций

Исследователям удалось манипулировать ключевыми словами и ухудшать выходные данные с помощью всего сотни токсичных записей, и они обнаружили, что большие модели менее стабильны и более уязвимы для отравления. Они также обнаружили, что предотвращение этих атак крайне сложно, если вообще реалистично возможно.

Тот факт, что доводка уязвима для отравления, вызывает опасения, потому что доводка базовых моделей с использованием пользовательских данных, похоже, является общей стратегией интеграции языковых моделей в корпоративное программное обеспечение и пользовательские услуги.

Но это также должно вызвать беспокойство у всех, кто использует услуги OpenAI, потому что до недавнего времени OpenAI использовала запросы конечных пользователей для доводки своих моделей.

Учитывая, что мы знаем об отравлении моделей уже несколько лет и учитывая сильные стимулы, которые имеет SEO-сообщество для манипуляции результатами, вполне возможно, что злоумышленники отравляли ChatGPT в течение нескольких месяцев. Мы не знаем об этом, потому что OpenAI не рассказывает о своих процессах, о том, как они проверяют запросы, которые используют для обучения, как они проверяют свой обучающий набор данных или как они доводят ChatGPT. Их секретность означает, что мы не знаем, насколько безопасно управлялся ChatGPT.

Им также придется в какой-то момент обновить свой обучающий набор данных. Они не могут оставить свои модели застрявшими в 2021 году навсегда.

Как только они это сделают, мы имеем только их слово - обещания на пальце - что они достаточно хорошо отфильтровали манипуляции с ключевыми словами и другие атаки на обучающие данные, что, как

предположил исследователь ИИ Эль Махди Эль Мхамди, математически невозможно в статье, над которой он работал, пока был в Google.

Это означает, что продукт OpenAI и ChatGPT переоценен. Мы не знаем, есть ли у их продуктов серьезные дефекты или нет. Это означает, что OpenAI, как организация, вероятно, переоценена инвесторами.

Единственным рациональным вариантом для нас остается оценивать их, как если бы их продукты были дефектными и манипулированными.

Мы должны искать альтернативы.