Оглавление

ЧАСТЬ III. ПОНИМАНИЕ РЕАЛЬНЫХ УГРОЗ

ЧАСТЬ IV. ЗАЩИТА

ГЛАВА 2

Мотивация к атакам

Сегодня технологии на базе глубоких нейронных сетей уже прочно вошли в нашу жизнь. Например, такие виртуальные помощники, как Amazon Alexa, Apple Siri, Google Assistant и Microsoft Cortana, используют модели глубокого обучения для понимания речевых аудиоданных. Многие алгоритмы для обеспечения и контроля сетевых взаимодействий (таких как веб-поиск) основаны на глубоких нейронных сетях для распознавания тех данных, которыми они оперируют. Модели глубокого обучения все чаще используются в областях применения с высокими требованиями к безопасности, таких как автономные транспортные средства.

Многие технологии на базе ИИ получают данные непосредственно из физического мира (например, с камер) или используют цифровые представления данных, предназначенные для человека (например, изображения, загружаемые на сайты социальных сетей). Это потенциально может вести к проблемам, поскольку при обработке данных из ненадежных источников любая компьютерная система становится уязвимой к атакам. За созданием вредоносных входных данных, использующих эти уязвимости, могут стоять самые разные мотивы, однако их можно разделить на следующие основные категории.

- *Уклонение*. Сокрытие контента от автоматического цифрового анализа. Примеры см. в разделах «Обход веб-фильтров» на с. 39, «Камуфляж против видеонаблюдения» на с. 42 и «Личная конфиденциальность в Интернете» на с. 43.
- *Влияние*. Воздействие на автоматизированные решения для получения личной, коммерческой или организационной выгоды. Примеры см. в разделе «Репутация в Интернете и управление брендом» на с. 41.
- *Дезориентация*. Создание хаоса с целью дискредитировать или нарушить работу организации. Примеры см. в разделах «Дезориентация автономных транспортных средств» на с. 44 и «Устройства с голосовым управлением» на с. 46.

В данной главе приведено несколько примеров возможной мотивации для создания вредоносных образов. Это далеко не полный список, но он дает представление о характере и разнообразии типов угроз.

Обход веб-фильтров

Сегодня организации испытывают все большую необходимость управлять получаемым извне веб-контентом для блокирования контента, который может рассматриваться как оскорбительный или неприличный. Это особенно касается компаний — владельцев социальных сетей и электронных торговых площадок, бизнес-модель которых зависит от внешних данных. Кроме того, многие компании связаны правовыми обязательствами по отслеживанию оскорбительных материалов и недопущению их дальнейшего распространения.

Эти организации сталкиваются со сложными задачами, которые с каждым днем усложняются. Они просто не могут найти достаточное количество людей для того, чтобы отслеживать все данные, загружаемые на сайты, с необходимой скоростью и блокировать их при необходимости. На сайты соцсетей ежедневно загружаются миллиарды сообщений. Их содержание не имеет четкой структуры и трудно поддается фильтрации; они могут содержать изображения, звуковую, текстовую информацию с трудноуловимой разницей между оскорбительным и неоскорбительным или законным и незаконным контентом. Отслеживать и фильтровать этот контент по мере его загрузки на сайт с привлечением людей просто невозможно.

Очевидное решение — использовать интеллектуальные машины, которые бы отслеживали, фильтровали или как минимум сортировали данные, как показано на рис. 2.1. В основе таких решений лежат глубокие нейронные сети. Их можно научить распознавать эмоциональную окраску и оскорбления в тексте, они могут классифицировать содержимое изображений и даже определять действия, выполняемые на видео. Например, ГНС можно научить распознавать изображения с намеками на употребление наркотиков, что позволит отсортировать такие изображения для их дальнейшей проверки человеком.

Когда отдельный пользователь или группа пользователей хочет загрузить контент, который не соответствует политике веб-сайта, возникает необходимость обойти систему фильтрации или отслеживания таким образом, чтобы загружаемый контент доносил предназначенную для людей информацию.

При этом отслеживающей веб-контент организации требуется постоянно повышать точность алгоритмов, выявляющих оскорбительный, неприличный и незаконный контент, а также осуществлять перехват вредоносных входных данных. Злоумышленнику же требуется совершенствовать вредоносный веб-контент по мере улучшения системы мониторинга, чтобы избегать распознавания искусственным интеллектом, обеспечивая донесение того же семантического значения человеку.

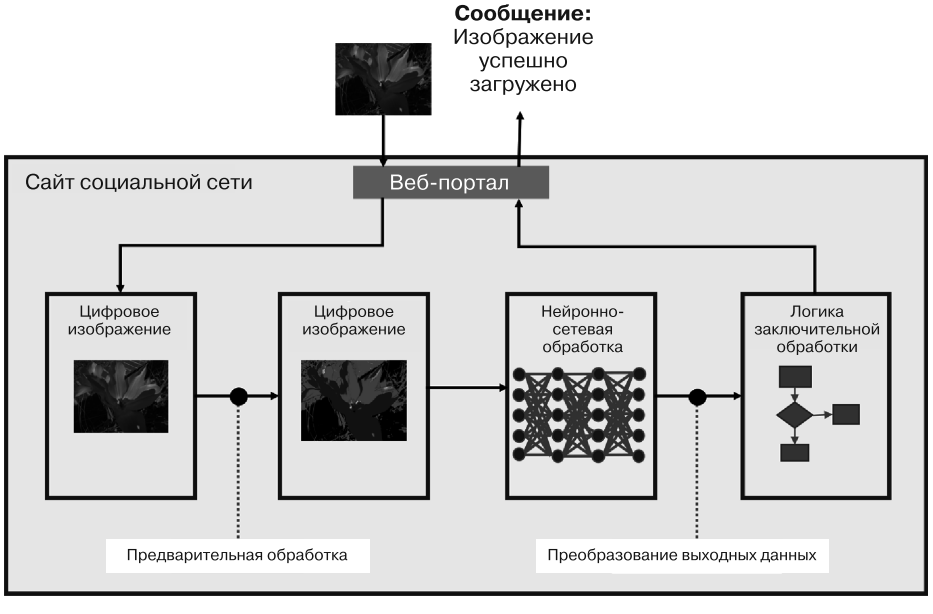

Рис. 2.1. Изображения, загружаемые на сайт социальной сети, могут подвергаться обработке и проверке искусственным интеллектом перед их добавлением на сайт

Злоумышленник может использовать еще один подход. Если обойти веб-фильтр загрузок невозможно, почему бы тогда просто не засыпать его множеством данных, обеспечивающих дезориентацию и дополнительные расходы для организации, обеспечивающей защиту? Принимаемое искусственным интеллектом решение о том, следует ли считать загружаемые данные «оскорбительными», обычно не является исключительно бинарным. Это скорее будет статистическая оценка вероятности с некоторым пороговым значением. Организации часто используют модерацию человеком для проверки изображений или данных, оценка которых близка к пороговому значению. Поэтому генерация большого количества неопасных данных, классифицируемых искусственным интеллектом как возможно опасные,

негативно скажется на оперативных возможностях организации и может снизить степень уверенности в точности результатов, выдаваемых искусственным интеллектом.

Если человеку трудно точно установить, *почему* данные были оценены как возможно нарушающие политику (поскольку они выглядят как неопасные), то большой наплыв данных потребует больших затрат времени и человеческих ресурсов — по сути, это атака типа «отказ в обслуживании».

Репутация в Интернете и управление брендом

Поисковые системы представляют собой сложные алгоритмы, которые не только решают, какие результаты следует отобразить в ответ на запрос, например, «кот на скейтборде», но и определяют последовательность этих результатов. Очевидно, что с коммерческой точки зрения лучше, если результат находится в начале списка. В силу этого компании мотивированы к тому, чтобы выяснить принцип действия алгоритмов поисковых систем и добиться того, чтобы их реклама отображалась на первой странице поисковой системы Google или Bing и была хорошо заметна при ее размещении на веб-страницах. Такая *оптимизация в поисковых системах* (search engine optimization, SEO), или просто поисковая оптимизация, является стандартной отраслевой практикой уже в течение многих лет. SEO играет ключевую роль в стратегии интернет-маркетинга.

Поисковые системы могут использовать специальных роботов для просеивания и индексирования страниц для отображения в результатах поиска на основе HTML-метаданных страницы, ее входящих ссылок и содержимого. Поисковые роботы представляют собой автоматизированные системы на основе ИИ, работающие без вмешательства человека. Поскольку информация заголовков легко поддается изменению, поисковые системы обычно больше полагаются на содержимое. Индексация на основе содержимого также делает возможным поиск с использованием менее распространенных ключевых слов, которые не всегда включаются в метаданные.

Характеристики страниц, основанные на их содержимом, особенно интересны в контексте вредоносных образов. Поскольку обновление изображений веб-сайта может повлиять на его позицию в результатах поисковой системы, у компании, желающей повысить видимость сайта для целевой аудитории, возникает мотив использовать вредоносные искажения или заплатки для изменения или подкрепления категории изображений без изменения восприятия их человеком.

Существует и более злонамеренный вариант атаки: стремясь подорвать доверие к организации или отдельному человеку, злоумышленник может сгенерировать вредоносные изображения, ведущие к ошибочной ассоциации объекта атаки с чем-то вредящим его репутации. Так, например, вредоносные изображения шоколадного батончика могут быть ошибочно классифицированы поисковой системой как «яд» и включены в число результатов, выдаваемых при поиске изображений ядов. Даже такая неявная ассоциация может серьезно сказаться на восприятии бренда потребителями.

Камуфляж против видеонаблюдения

Камеры видеонаблюдения получают свои данные из физического мира, что заставляет взглянуть на вредоносные входные данные с совершенно иной точки зрения по сравнению с предыдущими примерами. В данном случае цифровой контент генерируется на основе данных воспринимающего устройства (камеры) и не может изменяться посторонними лицами, которые не являются сотрудниками организации[1].

Хотя цифровые данные видеонаблюдения (видеозаписи или неподвижные кадры) часто по-прежнему отслеживаются людьми, эта задача становится все более трудноосуществимой из-за возрастания объема информации и затрат времени на ее обработку. В большинстве случаев данные видеонаблюдения можно не отслеживать активно в режиме реального времени, а анализировать позднее в медленном режиме (как, например, при расследовании преступлений). Организации все чаще прибегают к автоматизированным методам отслеживания или к сортировке данных с камер видеонаблюдения с использованием технологий на базе ИИ, например, для автоматического распознавания в данных видеонаблюдения конкретных лиц или транспортных средств с выдачей соответствующих оповещений.

Не нужно много фантазии, чтобы вообразить сценарии, в которых у злоумышленника может возникнуть желание обхитрить систему. Цель злоумышленника может состоять в том, чтобы создать своего рода «плащ-невидимку», способный обмануть ИИ, не привлекая лишнего внимания человека. При этом обычно достаточно просто не допустить активации (генерации) искусственным интеллектом тревожного сигнала, влекущего за собой тщательную проверку изображения или видеозаписи человеком. Например, злоумышленник может постараться обмануть систему распознавания лиц в режиме реального

[1] Очевидно, что это не так в случае более серьезных нарушений системы безопасности, когда посторонние лица получают доступ к внутренним данным организации.

времени, используемую службой безопасности аэропорта. Точно так же недопущение распознавания подозрительной активности системой обнаружения угроз в режиме реального времени дает злоумышленнику больше возможностей для совершения противоправных действий. При рассмотрении данных не в режиме реального времени может выявиться, что камеры видеонаблюдения зафиксировали информацию, относящуюся к преступлению, такую как лица преступников, номерные знаки и т. д., которую можно использовать для поиска на основе свидетельских показаний после совершения преступления. Скрыв эту информацию от ИИ, преступник может сделать выявление преступления менее вероятным.

Конечно, в основе такого обмана могут лежать не только криминальные мотивы. В мире, где отслеживается все и вся, люди стремятся закамуфлировать заметные черты лица от ИИ, чтобы сохранить личную конфиденциальность. Для достижения этой цели может использоваться, в частности, нейтральная одежда или макияж, что позволит привести правдоподобное отрицание против обвинений в намеренном обмане системы видеонаблюдения, объясняя ее неправильное функционирование просто как сбой в работе[1].

Существует еще один интересный сценарий: что, если злоумышленник внесет в реальном мире физические изменения, которые не будут казаться опасными и вредоносными для человека, но заставят систему видеонаблюдения поднять ложную тревогу? Это позволит злоумышленнику отвлечь ресурсы организации в ложном направлении, чтобы совершить реальное преступление в каком-либо другом месте.

Личная конфиденциальность в Интернете

Многие платформы социальных сетей для улучшения взаимодействия с пользователем извлекают информацию из загружаемых на сайт изображений. Так, например, Facebook регулярно извлекает и идентифицирует лица на изображениях для более эффективной разметки изображений, выполнения поиска и рассылки уведомлений.

Опять же пользователь, желающий сохранить личную конфиденциальность, может изменить изображения таким образом, чтобы затруднить распознавание лиц для использующегося платформой искусственного интеллекта. Для камуфлирования лиц от ИИ могут вноситься такие изменения, как наложение вредоносных заплаток на край изображения.

---

1 *Sharif et al.* Accessorize to a Crime.

Дезориентация автономных транспортных средств

Широко известным примером использования ИИ являются автономные транспортные средства, которые относятся к системам с высокими требованиями к безопасности. Такие транспортные средства функционируют в неупорядоченном, неограниченном и постоянно меняющемся физическом мире. Уязвимость к вредоносным входным данным может при этом привести к катастрофическим последствиям.

Автономными могут быть не только автомобили. Сегодня автономность получает все более широкое распространение на море, в воздухе и под водой. Автономные транспортные средства также используются в ограниченных и закрытых окружениях, таких как производственные помещения, для выполнения основных или, возможно, опасных задач. Даже при таком ограниченном окружении существует риск получения с камеры вредоносных входных данных, внесенных сотрудником организации (внутренняя угроза) или лицом, получившим доступ к рабочей зоне системы. В то же время не следует забывать о том, что, как правило, автономные транспортные средства контролируют физическое окружение, руководствуясь не только данными воспринимающего устройства. Большинство автономных систем получает информацию из нескольких источников, которые могут включать:

- *внебортовые данные*. Большинство автономных транспортных средств руководствуется данными, получаемыми из одного или нескольких внебортовых централизованных источников[1]. Внебортовые данные включают в себя сравнительно статичную информацию (карты и скоростные ограничения), централизованно собираемые динамические данные (например, сведения об интенсивности дорожного движения) и данные, относящиеся к конкретному транспортному средству (например, его GPS-координаты). Все эти типы источников данных уже используются в таких приложениях для GPS-навигации, как Waze, Google Maps и HERE WeGo.

  В других областях применения могут использоваться иные виды внебортовых данных. Например, в сфере морских перевозок широко применяется автоматическое отслеживание координат морских судов по данным *автоматических идентификационных систем* (АИС). С помощью этих систем корабли регулярно передают в режиме реального времени свои идентификационные данные и координаты, что позволяет органам морской власти следить за их передвижением;

[1] В таких ограниченных окружениях, как производственные или жилые помещения, автономные транспортные средства иногда руководствуются исключительно данными бортовых датчиков.

- *данные бортовых датчиков.* Автономное транспортное средство также может руководствоваться данными таких бортовых датчиков, как камеры, датчики расстояния, акселерометры и гироскопические датчики (для выявления позиционного вращения). Эти данные играют решающую роль в предоставлении информации об изменениях в непосредственной близости от транспортного средства, например, при подаче тревожного сигнала в режиме реального времени или при каких-либо неожиданных событиях.

В некоторых случаях автономное транспортное средство принимает решение лишь на основе данных датчиков. Например, автономное транспортное средство может корректировать свое положение на дороге, руководствуясь исключительно данными датчика, как показано на рис. 2.2. Такой сценарий может представлять существенную угрозу безопасности, поскольку генерируемая информация потенциально ненадежна.

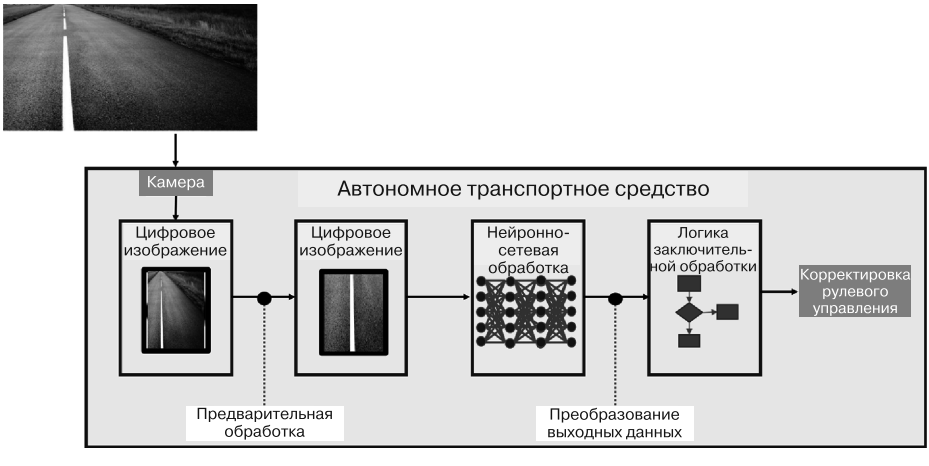

Рис. 2.2. Автономное транспортное средство может корректировать свое положение на дороге на основе данных камеры

На практике автономные транспортные средства обычно руководствуются информацией, получаемой из нескольких источников данных, и действуют с некоторой долей перестраховки. Если дорожный знак остановки STOP будет вредоносным образом «превращен» в разрешающий движение знак преимущества движения, это вряд ли обманет транспортное средство, которому также доступны централизованные данные о регулировании дорожного движения (сведения о скоростных ограничениях, перекрестках и местах, где нужно остановиться или уступить дорогу). С большей степенью вероятности злоумышленники попытаются «сыграть» на этой осмотрительности

транспортного средства — например, они могут парализовать работу дорожной сети, усеяв дорогу множеством неопасных с виду наклеек, ошибочно интерпретируемых как опасные объекты.

Устройства с голосовым управлением

Голосовое управление предоставляет естественный бесконтактный способ управления многими аспектами нашей жизни. С его помощью можно выполнять широкий круг задач — от управления мультимедийными устройствами и домашней автоматизации до поиска товаров и совершения покупок в Интернете. Такие устройства с голосовым управлением, как смартфоны, планшеты и голосовые помощники, постепенно занимают прочное место в нашем доме. Лежащая в основе этих устройств обработка речи осуществляется с помощью продвинутых технологий на базе ГНС и отличается высоким уровнем точности. На рис. 2.3 показано, как может выглядеть простейшая последовательность обработки для устройства с голосовым управлением.

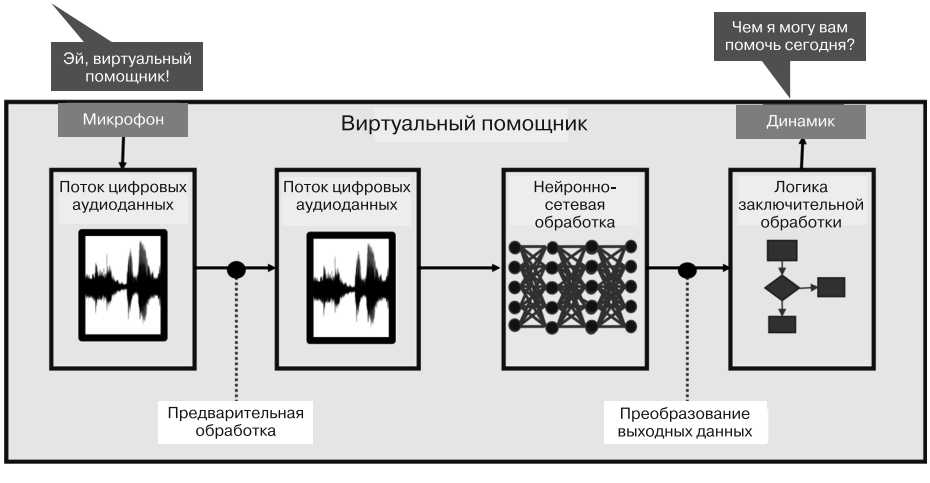

Рис. 2.3. Виртуальный помощник использует ИИ для обработки речевых данных и выдачи соответствующего ответа

Поток аудиоданных может поступать в дом по каналам радио- и телевещания или в виде сетевого контента. Незаметно для пользователя в эти аудиоданные может быть включен вредоносный контент, который, к примеру, будет давать голосовому помощнику указание увеличить громкость проигрываемой песни. При всей незначительности таких лишь немного раздражающих отклонений

в работе голосового помощника они тем не менее могут подорвать доверие к нему со стороны пользователей. Помощник, который ведет себя непредсказуемо, будет раздражать пользователей и в конечном итоге вызовет у них отвращение. А если устройство утратит доверие пользователей в домашнем окружении, ему будет очень сложно заслужить его снова. Потенциально скрытые команды могут оказаться и не столь безобидными — например, они могут самопроизвольно отправить СМС или сделать запись в социальной сети, изменить настройки устройства, выполнить переход по вредоносной ссылке или изменить настройки домашней системы безопасности.

Для выполнения функций, которым необходим повышенный уровень безопасности, голосовые помощники требуют дополнительных мер безопасности, чтобы не допустить их случайного или умышленного неправильного использования. Помощник может спросить: «Вы *действительно* хотите приобрести экземпляр книги "Надежность нейронных сетей"?» После этого будет ожидать подтверждения покупки. Вы вряд ли скажете «да» в нужный момент, если эта покупка не была инициирована вами, хотя такое и не исключено. В то же время если есть возможность дать голосовому помощнику вредоносную *команду*, то также существует и возможность дать ему вредоносный *ответ*, конечно, при условии, что поблизости не будет никого, кто бы услышал запрос на подтверждение.

Важной особенностью здесь является то, что вредоносные образы не всегда действительно наносят вред — они могут использоваться в развлекательных целях как дополнительный способ передачи команд. С их помощью можно не только причинить вред, но и извлечь коммерческую выгоду.

Представьте, что вы запаслись попкорном и сели перед телевизором, чтобы еще раз посмотреть фильм «Сияние». Однако вы решили посмотреть не обычную версию этого фильма ужасов, а версию с «интегрированной поддержкой домашней автоматизации». Эта версия с расширенной мультимедийной поддержкой содержит вредоносные аудиоданные — скрытые сообщения для вашей системы домашней автоматизации, призванные расширить ощущения от просмотра. Так, в какой-то момент в доме может громко хлопнуть дверь, отключиться свет или, возможно, даже отопление (чтобы вы немного похолодели от ужаса)...

Что ж, на этой слегка тревожной ноте перейдем к следующей главе, где рассмотрены основные понятия ГНС.