Состязательные атаки для автономных транспортных средств

Д.Е. Намиот, В.П. Куприяновский, А.А. Пичугов

***Аннотация*—В данной статье рассматриваются состязательные атаки, направленные на модели машинного (глубокого) обучения, используемых в автономных транспортных средствах. Системы Искусственного интеллекта (машинного обучения) играют определяющую роль в функционировании беспилотных автомобилей. В то же самое время, все системы машинного обучения подвержены так называемым состязательным атакам, когда атакующий сознательно модифицирует данные так, чтобы обмануть алгоритмы работы таких систем, затруднить их работу (понизить качество работы) или добиться желаемого атакующим поведения. Состязательные атаки представляют собой большую проблему для систем машинного обучения, особенно при использовании последних в критических областях, таких как автоматическое вождение. Состязательные атаки представляют собой проблему для функционального тестирования – есть данные, на которых система работает неверно (вообще не работает, работает с низким качеством). Для систем автономного транспорта такие атаки могут осуществляться в физической форме, когда модифицируются реальные объекты, захватываемые сенсорами транспортного средства, создаются фиктивные объекты и т.д. В настоящей статье приводится обзор состязательных атак на беспилотные транспортные средства, где основное внимание уделено именно физическим атакам.**

***Ключевые слова*—машинное обучение, глубокое обучение, состязательные атаки.**

I. Введение

Модели машинного обучения подвержены состязательным атакам. Состязательные атаки – это модификации данных на разных этапах стандартного конвейера машинного обучения, которые либо препятствуют корректной работе модели, либо заставляют ее работать нужным атакующему способом. Атаки могут также принимать форму специальных запросов, которые призваны выявить непубличную информацию о тренировочных данных, либо восстановить алгоритм работы модели. NIST, в последнем варианте своего классификатора состязательных атак [1], выделяет атаки уклонения, отравления и атаки на приватные данные (их уже 5 типов). Для генеративных систем есть еще атаки злоупотребления [2]. Этими, основными способами воздействия на модели через данные, дело не ограничивается [3]. Предобученная модель, с практической точки зрения, представляет собой файл, который может быть инфицирован вредоносным контентом [4]; модель исполняется в некоторой программной среде, которая может воздействовать на модель (например, слепая атака - модификация функции подсчета потерь в варианте стандартного фреймворка и т.п.). Соответственно, состязательные атаки представляют собой очень серьезную проблему для практического использования систем машинного обучения (систем искусственного интеллекта). Наибольшую проблему это представляет для так называемых критических применений (авионика, автоматическое вождение и т.п.). Кратко, состязательные атаки препятствуют функциональному тестированию, которое является обязательной частью сертификации программных систем для критических применений [5]. Состязательные примеры – это и есть примеры, на которых модель (алгоритм) работает неверно. Работы, посвященные робастности моделей машинного обучения, которые во многих странах инициируются на государственном уровне, как раз и посвящены, в том числе, проблеме состязательных атак [6,7].

Атаки уклонения – это состязательные атаки, которые происходят на этапе использования (выполнения) модели [8]. Для моделей, работающих в автономных транспортных средствах – во время эксплуатации (движения). В данном случае (автономные транспортные средства) – это самый реалистичный сценарий атаки.

Атаки уклонения принято разделять на цифровые и физические [3]. В первом случае модифицируются цифровые данные, во втором – как-то изменяется физическая среда (окружение). Последнее представляет собой самый реалистичный сценарий атаки для атак уклонения. Важно, что работающая модель никоим образом не может предотвратить такого рода воздействия (не может запретить атакующим манипуляции в окружающей среде).

Другие формы состязательных атак также возможны. Если модели, например, обучаются на публичном датасете, то нет гарантий, что там нет отравленных

Статья получена 19 мая 2024.
Д.Е. Намиот – МГУ имени М.В. Ломоносова (e-mail: dnamiot@gmail.com).
В.П. Куприяновский – РУТ(МИИТ) (e-mail: v.kupriyanovsky@rut.digital).
А.А. Пичугов – ООО Матричные системы (email: info@matrixcompany.ru)

(специально модифицированных) данных [9]. Если разработчики использовали предобученные модели, то возможны трояны. И, конечно, модели машинного обучения в автономных транспортных средствах должны работать на доверенных платформах [10].

Относительно атак, которые пытаются получить приватные данные от моделей, здесь нужно смотреть на модель применения. Все такие атаки связаны с множественными запросами к моделям [11]. Представляется, что поддержка сервера с публичным API – это не самый реалистичный вариант использования модели машинного обучения в транспортном средстве. Такого рода атаки больше касаются MLaaS систем [12].

И, конечно, нужно отметить, что проблемы кибербезопасности автономных транспортных средств не ограничиваются состязательными атаками на модели машинного обучения (фактически, на систему управления). По крайней мере, есть еще связь (V2V, V2I), со своими проблемами [13].

К настоящему времени зафиксировано несколько аварийных происшествий, вызванных ошибочным или ненормальным поведением систем с автономным вождением (АВ), в неопределенных и сложных условиях. Есть несколько конкретных примеров заметных сбоев. Например, Tesla AV, работающая в режиме «Автопилот», привела к гибели во Флориде [14]. Точно так же Uber с системой автономного вождения столкнулся с пешеходом в штате Аризона [15]. Есть информация, что Uber участвовал в 37 других авариях до аварии со смертельным исходом в 2018 году [16]. За автопилотом Tesla числится 13 смертельных аварий [17].

Кроме того, было несколько случаев, когда хакеры выявляли угрозы кибербезопасности в расширенных функциях помощи водителю, доступных в легковых автомобилях. Например, исследователи из Keen Security Labs в Китае продемонстрировали пару эксплойтов через систему камер в Tesla Model S [18]. Последняя работа является примером атак на саму систему АВ, в которой модель (модели) машинного (глубокого) обучения является лишь частью [19-21]. В данной же работы мы концентрируемся именно на состязательных атаках на системы машинного обучения, которые составляют ядро управления АВ.

Оставшаяся часть статьи структурирована следующим образом. В разделе II мы рассматриваем примеры физических атак. Раздел III посвящен атакам на системы автоматического вождения.

II. Физические атаки

Физические атаки (атаки в физическом мире) – это реальные модификации окружения, в котором работает модель. Цель этих модификаций – изменить восприятие окружения моделью, так, чтобы это изменило ее работу (ухудшило качество, заставило принимать неверные/нужные атакующему решения по классификации и т.п.).

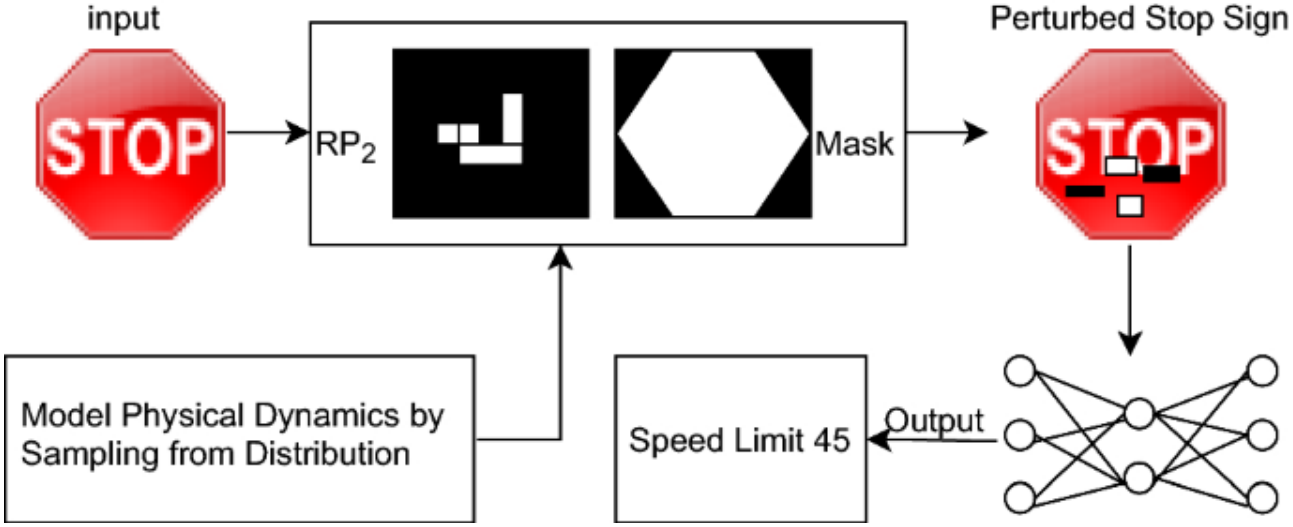

Рис.1. Физическая атака патчами [22]

На рисунке 1 [22] представлен пример такой атаки. Здесь вычисляется патч для дорожного знака, который меняет его классификацию.

Физические атаки начинались с работы 2018 года [23], в которой авторы построили состязательно возмущенное изображение, напечатали его на бумаге, а затем сняли на камеру мобильного телефона. Полученное изображение продолжало “обманывать” нейронную сеть. На рисунке 2 изображена общая схема подобных атак:

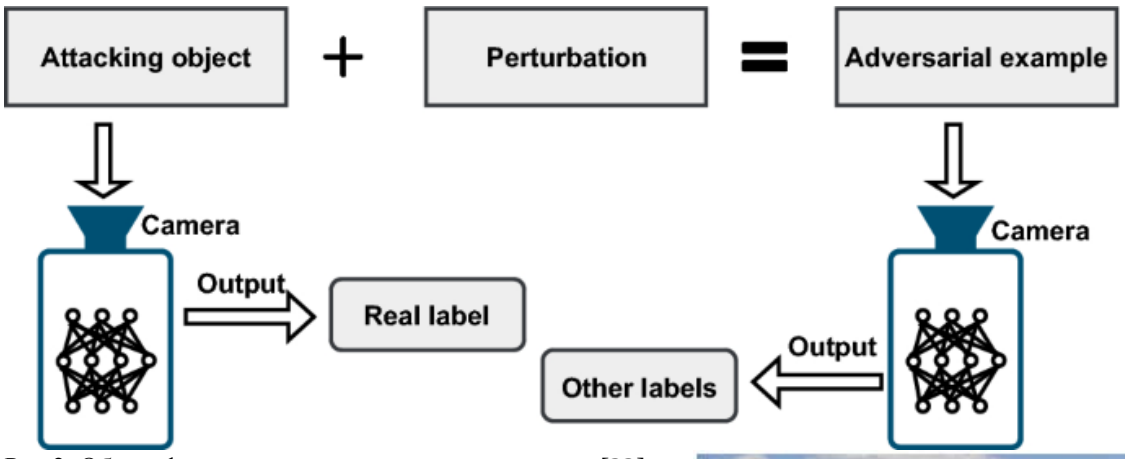

Рис.2. Общая форма атак на систему распознавания [22]

Объект атаки с добавленным возмущением захватывается датчиком (например, камерой, LiDAR), а затем передается в модель для распознавания, и метка оказывается другой.

Из этого проистекают главные моменты такого рода атак:

- Как сформировать возмущение (создать фиктивный объект)
- Как доставить новую информацию до датчиков объекта (датчиков АВ в данном случае)

Некоторые креативные примеры физических атак представлены ниже.
На рисунке 3 представлена состязательное вязание (вышивка). Получающаяся картинка соответствует характеристикам (фичам), которые получаются при распознавании лиц. Соответственно, система распознавания “видит” множество лиц (зеленые прямоугольники).

Рис. 3. Состязательная одежда [24].

На рисунке 4 представлен “транспортный” пример [3]. Здесь атакующие воспользовались тем фактом, что система распознавания дорожных знаков в автопилоте работает хорошо, распознает знаки устойчиво, но не проверяет (не умеет определять) контекст, в котором этот знак появляется (рис. 4).

Рис. 4. Знак дорожного движения в произвольном месте [25]

Размещение “знака” дорожного движения с помощью дрона и проектора на столбе успешно распознается автопилотом (рис. 5).

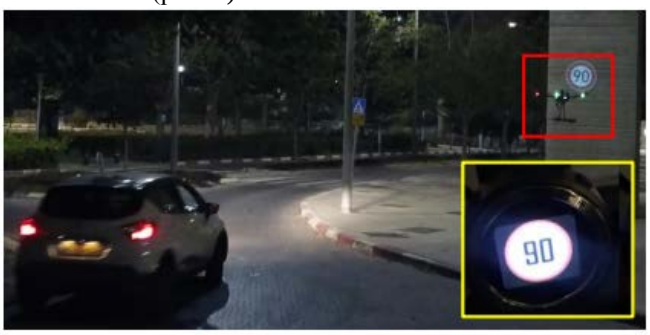

Рис. 5. Проецирование знака [25]. В желтом прямоугольнике – снимок экрана автопилота.

В данном случае, из соображений безопасности, использовался фиктивный знак ограничения скорости движения. При его успешном распознавании автомобиль просто замедлил бы движение. Но, с таким же успехом, распознается и знак “Только направо”, также размещенный вне контекста. Это открывает путь к опасным физическим атакам: АВ транспорт движется в крайней правой полосе, в полосу движения перестраивается грузовик (фургон) с нанесенным на заднем борту знаком “Движение направо” …

На рисунке 6 автопилот также успешно распознает “препятствие” на дороге.

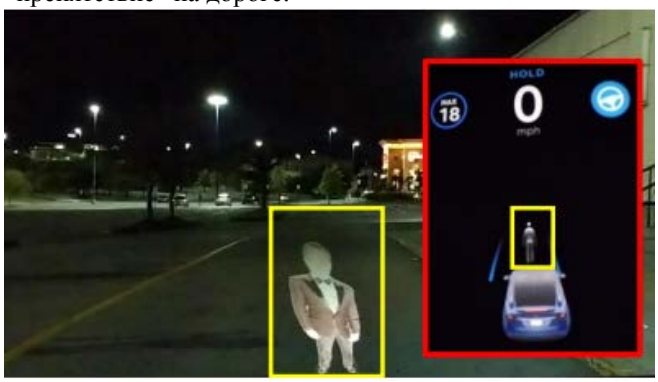

Рис. 6. “Препятствие” на дороге [25]. В красном прямоугольнике – снимок экрана автопилота.

Этот пример демонстрирует главный креативный момент в физических атаках – как подать нужные данные в регистратор. В данном случае использовался проектор для создания фиктивных знаков.

В целом, проблема контекста достаточно серьезная при распознавании изображений. Вот в статье [26] собраны отзывы на испытания трамвая с автоматическим движением в Санкт-Петербурге. Там упоминаются случаи, когда система реагирует на “посторонние” светофоры, движения на соседних линиях и т.п.:” Водители жаловались, что вагон оттормаживается из-за каждого перестроения соседних автомобилей на рельсы, считывает как препятствия столбы у дороги и другие объекты, на миг попадающие в створ движения, например, при повороте… Система плохо видит светофоры, не различает «свои» и чужие, ловит красный сигнал со следующего перекрестка и т.п.” В этом случае могло бы помочь сетевое взаимодействие транспортного средства с другими (V2V) или с инфраструктурой (V2I). Но тогда были бы другие состязательные атаки (DDOS, например) [27].

Важным моментом для атак в реальном мире является тот факт, что модель машинного обучения не может никаким образом предотвратить их появление. Транспортное средство с системой АВ никак не может предотвратить манипуляции атакующих с окружением. Другим важным моментом является то, что физические модификации должны быть устойчивы к разным режимам их восприятия. Например, если мы говорим об упомянутых выше патчах (накладных картинках), то можно отметить, что в реальном мире камера будет снимать их под разными углами, при разном освещении, в дождь, туман, загрязненными и т.п. При физической печати состязательного изображения (патча) количество цветов и переходы между цветами будут естественным образом (возможности печатной машины) ограничены, по сравнению с манипуляциями с изображением в цифровом мире.

В работе [22] приводится следующая таксономия для практических состязательных атак в физическом мире:

- Распознавание изображений (атаки на определение дорожных знаков, атаки на камеры, атаки на системы определения полосы движения)
- Распознавание объектов (атаки на системы определения пешеходов, атаки на лидары, инфракрасные атаки)
- Неразличимые на слух голосовые команды

III Состязательные атаки на АВ

В работе [28] описывается модель построения рекламных билбордов (отдельно стоящая металлическая конструкция на опоре с рекламным щитом), которые заставляют системы управления изменять направление движения (рис.7). Стрелки указывают измененное направление движения. Это атака белого ящика, авторы предполагают, что модель управления используют сверточную сеть. Это важный момент – большая часть из рассматриваемых физических атак – это именно атаки белого ящика. CNN действительно широко используется в системах управления автономного транспорта. Авторы использовали для тестирования конкретные системы [29, 30]. С практической точки зрения атака белого ящика означает, что атакующему необходимо будет получить информацию о системе управления АВ или ее тренировочных данных, чтобы построить теневую модель. Впрочем, наборы данных для обучения в данном случае должны быть похожи (по сути, они будут соответствовать стилю езды – технике прохождения поворотов).

Цель – генерация именно печатаемого рекламного плаката. Для этого каждый пиксель вычисленного изображения должен быть напечатан.

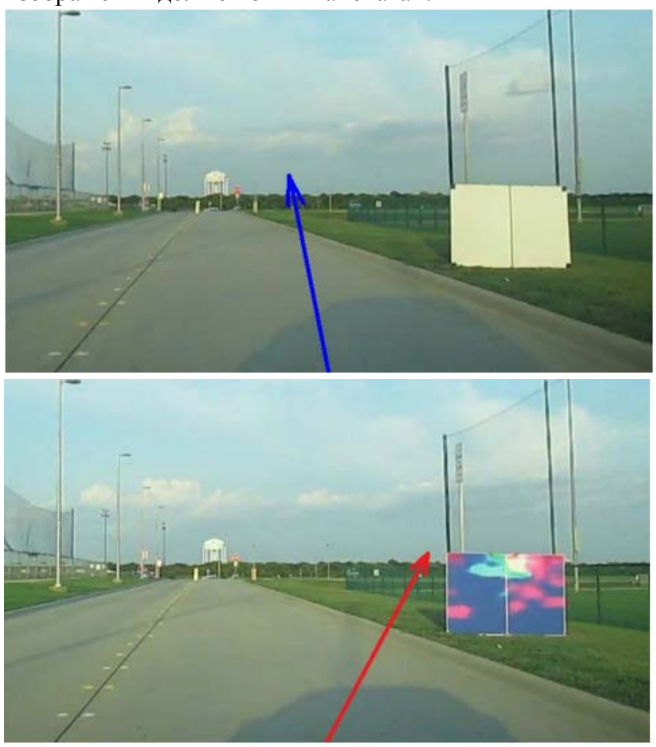

Рис.7. Состязательные билборды [28].

Пусть $P \subset [0, 1]^3$ — набор распечатываемых пикселей RGB. Авторы определили показатель непечатности (*NPS*), который отражает максимальное расстояние между этим пикселем и любым пикселем в P. Большее значение NPS будет означать меньшую вероятность точной распечатки соответствующего пикселя. Таким образом, алгоритм стремится минимизировать NPS в рамках оптимизации. Для всего изображения NPS определяется как сумма NPS всех пикселей.

Предложенный алгоритм состоит в том, что на видеорегистратор записываются проезды рекламного щита, заполненного одним цветом. Выделяются только углы изображения, для того, чтобы модифицировать пиксели в заданных границах. Далее полученные видео разбиваются на отдельные кадры и итеративно модифицируются пиксели, с расчетом градиента решения модели по выставлению угла поворота (вот здесь и нужен белый ящик). Цель – максимзировать изменение угла поворота для всех последовательных кадров.

Код решения доступен [31]. В реальных условиях атака DeepBillboard может довести ошибку рулевого управления беспилотным транспортным средством до 26,44°.

В работе [32] такая же задача решается для виде-билбордов, где можно в реальном времени демонстрировать состязательные изображения. Отметим, что в таком случае нет проблем с физической распечаткой символов, но остаются проблемы с углом зрения, освещенностью и т.п. Чтобы дополнить эту схему атаки отметим, что билборды, вообще говоря, бывают и мобильными (размещенными на автомобилях – рис. 8). Это делает такую схему атаки весьма опасной.

Рис. 8. Мобильный билборд [33]

Система помощи при удержании полосы движения (Lane-Keeping Assistance System - LKAS) — это технология автоматизации вождения уровня 2, которая автоматически управляет транспортным средством так, чтобы оно оставалось в пределах текущей полосы движения [34]. Благодаря удобству для водителей, система широко доступна в различных моделях автомобилей, таких как Honda Civic, Toyota Prius, Nissan Cima, Volvo XC90 и т.д. В основе работы – нейронная сеть (DNN), которая также становится объектом атак. Типичная работа [34] описывает физическую атаку патчами (рис. 9): имитатор загрязнения частично перекрывает полотно дороги (но не саму разметку!). Собственно линия разметки не перекрывается из желания сохранить атаку тайной. Этой же цели следует серый цвет патча (имитация грязи на дороге). В работе описывается атака белого ящика – использовалась известная реализация LKAS – OpenPilot [35].

Рис.9 Имитация загрязнения [34]

Для водителя (камеры) это выглядит так (рис. 10):

Рис. 10. “Загрязнение” в полосе [34].

При вычислении патча брались последовательные кадры дорожного полотна, и на них накладывались патчи. Код решения доступен [36]. Общая схема решения представлена на рисунке 11. Поскольку размеры и угол видимости патча на каждом кадре будут разные, то значения пересчитывались на BEV (bird’s eye view – вид с высоты птичьего полета). Опубликованные тесты показывают “увод” автомобиля с полосы за 0.9 сек (в зависимости от размера патча).

Интересно, что в работе [42] рассматривается физическая бэкдор атака на систему определения полосы движения, где специальным образом отравляются изображения, на которых тренируется система распознавания в автономном транспортном средстве.

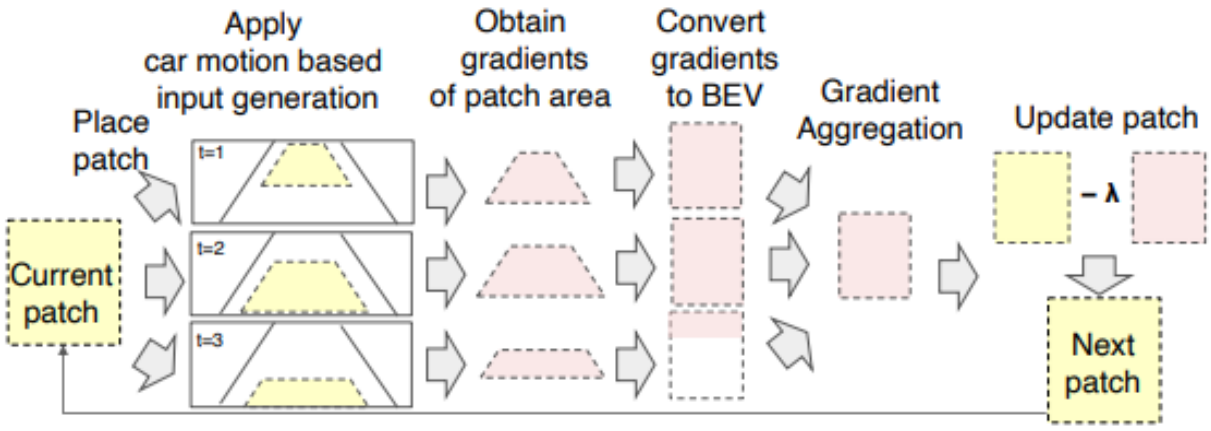

Рис. 11. Оптимизационная задача для вычисления патча [34].

Идея с использованием состязательных атак на системы распознавания дорожных знаков возникла практически сразу, как появились состязательные атаки. Из знаковых работ в этой области отметим работу [37], где авторы исследовали, в том числе, и атаки черного ящика, когда у атакующего нет информации о модели распознавания. При этом авторы сделали одно очень важное, на наш взгляд, дополнение: исследовали состязательные атаки с ограничениями на модификации, когда модифицируются только наиболее значимые для распознавания области. Это может понизить успешность атак, но зато очень сильно повысит их скрытность. Результат показан на рисунке 12, сделанном по данным работы [37]. Первое и второе изображения – результат атаки черного ящика, которое меняет распознавание знака со Stop на ограничение скорости в 60 км/ч, но во втором случае – модификации ограничены.

Рис. 12. Ограниченная атака черного ящика [37]. Очевидно, что в правом изображении гораздо труднее заподозрить атаку.

В работе [38] решается, в некотором роде, обратная задача – как добиться того, чтобы рекламные плакаты распознавались как ложные дорожные знаки (рис. 13).

Рис. 13. Слева – чистый логотип, справа – знак Стоп. [38].

Это атака белого ящика, исходный код – доступен. Это так называемая OOD (out-of-distribution) атака. Атакующий не модифицирует чистое атакуемое изображение, а начинает с произвольного образа.

Авторы продолжили эту тему в работе [39], где предложили так называемую атаку лентикулярной печати. Лентикуля́рная печа́ть (линзора́стровая печа́ть) - технология печати изображений, в которой массив из плоско-выпуклых цилиндрических собирающих линз (лентикулярный растр) используется для создания иллюзии глубины пространства и многоракурсности или смены изображения при просмотре под разными углами. Один из способов автостереоскопии, или безочковой сепарации изображений стереопары в 3D-фотографии [40]. Примеры – “переливающиеся” значки, стереооткрытки и т.п. В случае атак на автономный транспорт, эта технология позволяет менять “изображение” знака в зависимости от угла, под которым он виден камере автономного транспортного средства. Предложенный процесс атаки был расширен в работе [41], где авторы в режиме атаки черного ящика модифицировали знаки ограничения скорости. Отметим, что атаки черного ящика в данном случае заключались в том, что произвольно выбиралась модель, на ней строились атаки (в режиме белого ящика), которые затем тестировались на других моделях, информация о которых была недоступна при построении атаки. Результаты реальных испытаний (автополигон) показали очень большой процент успешных обманов.

Отметим, что о цифровых атаках, когда прямо модифицируется изображение камер транспортного

средства, также говорят в статьях (например, [43]), но речь здесь идет о взломе системы управления и доступе к CAN-шине (внутренним коммуникациям). Поэтому эта тема вне рассмотрения данной статьи.

Физические атаки существуют и для моделей детектирования объектов, которые используются в автономных транспортных системах. Например, работы [44, 45] описывают физические состязательные атаки на модели YOLO и R-CNN. Если в работе [46] авторы проверяли, как нанесение камуфляжного рисунка методом аэрографии влияет на определение марки автомобиля, то в работе [47] авторы задались задачей подобрать такой камуфляж, который полностью исключит детекцию объекта (автомобиля).

В статье [48] описывается робастная физическая атака на детектор дорожных знаков, использующий YOLO. В отличие от ранее описанных атак, здесь речь идет не об ошибочной классификации, а о том, чтобы система распознавания вместо дорожного знака видела совсем другой объект. Цель злоумышленника — не дать детектору объектов обнаружить целевой объект. Чтобы добиться этого, состязательное возмущение должно гарантировать, что вероятность появления целевого объекта в любой ограничивающей рамке меньше порога обнаружения. Авторы используют модифицированный подход R_2 [49], который состоит в том, чтобы выполнить выборку из распределения, имитирующего физические возмущения объекта (например, расстояние до него и угол обзора), и найти возмущение, которое максимизирует вероятность неправильной классификации при этом распределении.

Отметим, среди других, работу [50], где авторы предложили оригинальную схемы атаки черного ящика на систему распознавания дорожных знаков: вместо того, чтобы добиваться максимизации потерь в предсказании правильного класса, модель занимается минимизацией потери для неправильного класса, предсказанного моделью с наибольшей вероятностью. Итог такой атаки - выравнивание распределения вероятности неверных классов и создания состязательных данных с наименьшей уверенностью в истинном классе.

Лидары (LIDAR - Light Detection and Ranging) также не остались без внимания атакующих [51]. Датчик LiDAR излучает короткие импульсы инфракрасного света. Инфракрасный свет отражается от объектов, попадающих в поле зрения датчика. Датчик измеряет время, за которое свет возвращается обратно. На основе этого времени датчик вычисляет расстояние до объекта. Обработка (анализ) отраженных сигналов выполняется с помощью модели машинного обучения. Рисунок 14 показывает пайплайн обработки

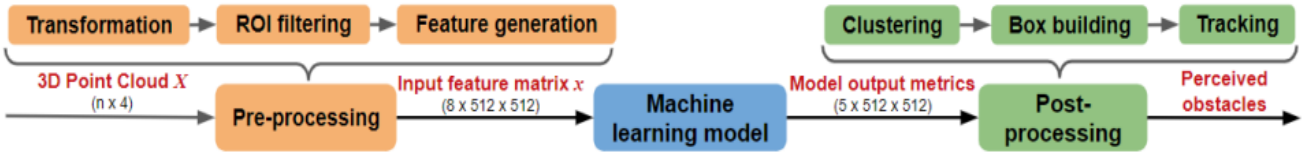

Рис. 14. Обработка данных лидара [51]

Соответственно, эта модель и становится объектом атаки. Физическая атака заключается в спуфинге – размещении ложных источников сигнала, которые “обстреливают” сенсоры транспортного средства. Чтобы вызвать семантически важные последствия для безопасности в настройках АВ, авторы поставили целью атаки обман восприятия на основе LiDAR, заставив его воспринимать ложные препятствия перед атакуемым АВ, чтобы злонамеренно изменить его решения о вождении. Атака нацелена на искусственные препятствия, расположенные спереди, то есть на те, которые находятся на близком расстоянии от передней части атакуемого АВ, поскольку они имеют самый высокий потенциал инициировать немедленные ошибочные решения по вождению. В работе определяются препятствия, расположенные спереди и на расстоянии около 5 метров от атакуемого АВ.

Существует несколько возможных сценариев проведения такой атаки. Во-первых, злоумышленник может разместить атакующее устройство на обочине дороги, чтобы стрелять вредоносными лазерными импульсами по проезжающим мимо автомобилям. Во-вторых, злоумышленник может управлять атакуемым транспортным средством в непосредственной близости от жертвы, например, двигаясь по той же полосе или по соседним полосам движения. Для проведения атаки атакующая машина оснащена атакующим устройством, которое посылает лазерные импульсы в LiDAR атакуемого АВ. Для выполнения лазерного прицеливания в этих сценариях злоумышленник может использовать обнаружение и отслеживание объектов с помощью камеры. В условиях АВ эти атаки являются скрытными, поскольку лазерные импульсы невидимы, а устройства лазерной стрельбы относительно малы по размеру.

Спуфинговые атаки на камеры транспортных средств довольно легко воспроизводимы и несут вполне очевидные последствия. Это представлено на рисунке 15 из обзора [16]: неожиданное торможение перед “фиктивным” объектом вызывает столкновение со сзади идущим автомобилем

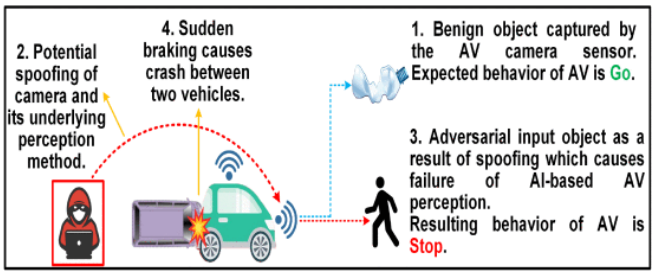

Рис.15. Спуфинг камеры [16].

Вопросы защиты от рассмотренных атак требуют, безусловно, отдельной работы. Здесь и уменьшение количества признаков, энкодеры с шумоподавлением (DAE – denoising autoencoder), карты глубины и т.п. Но общая схема с атаками на модели остается такой же, как и в других доменах – атаки опережают защиты. Сначала появляется атака, потом – какие-то способы смягчения ее последствий. Воспроизведение традиционной модели сертификации программного обеспечения для критических применений по отношению к моделям машинного обучения в беспилотных транспортных средствах пока не получается [52].

IV Заключение

В работе рассмотрены состязательные атаки для систем машинного обучения, используемых в автономных транспортных средствах. Как и другие состязательные атаки, рассмотренные примеры являются главным препятствием для использования систем машинного обучения (искусственного интеллекта) в критических применениях. При этом мы рассматривали только состязательные атаки на модели машинного (глубокого) обучения, которые являются ядром для беспилотных транспортных средств. Проблема кибербезопасности автономных транспортных средств, естественно, более широкая.

Физические атаки на модели машинного обучения в беспилотных средствах представлены очень широко и являются достаточно легко воспроизводимыми. Защита моделей посредством разного вида состязательных тренировок в любом случае не дает полной гарантии для неизвестных атак. Возможно, что решение той же проблемы распознавания дорожных знаков лежит в расширении (дополнении) визуального подхода. Например, помимо картинки, знак может обозначать себя как элемент (объект) сетевой инфраструктуры с уникальным ID, привязанным к местоположению и т.д.

Благодарности

Авторы благодарны сотрудникам кафедры ИБ факультета ВМК МГУ имени М.В. Ломоносова за обсуждение работы и ценные замечания. Работа продолжает линию публикаций, выполненных для поддержки магистерских программ факультета ВМК МГУ имени М.В. Ломоносова – Искусственный интеллект в кибербезопасности [53] и Кибербезопасность [54, 55].

Традиционно отметим другие работы В.П. Куприяновского и его многочисленных соавторов, которые и положили начало серии публикаций в журнале INJOIT о цифровой трансформации [56-58].

Библиография

[1] NIST AI 100-2 E2023 Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. https://csrc.nist.gov/pubs/ai/100/2/e2023/final Retrieved: May, 2024

[2] Shibli, Ashfak Md, Mir Mehedi A. Pritom, and Maanak Gupta. "AbuseGPT: Abuse of Generative AI ChatBots to Create Smishing Campaigns." arXiv preprint arXiv:2402.09728 (2024).

[3] Namiot, Dmitry. "Schemes of attacks on machine learning models." International Journal of Open Information Technologies 11.5 (2023): 68-86.

[4] Bidzhiev, T. M., and D. E. Namiot. "Attacks on Machine Learning Models Based on the PyTorch Framework." Avtomatika i telemehanika 3 (2024): 38-50.

[5] Namiot, Dmitry, and Eugene Ilyushin. "Trusted Artificial Intelligence Platforms: Certification and Audit." International Journal of Open Information Technologies 12.1 (2024): 43-60.

[6] Namiot, Dmitry, Eugene Ilyushin, and Ivan Chizhov. "Ongoing academic and industrial projects dedicated to robust machine learning." International Journal of Open Information Technologies 9.10 (2021): 35-46. (in Russian)

[7] Намиот, Д. Е., Е. А. Ильюшин, and И. В. Чижов. "Основания для работ по устойчивому машинному обучению." International Journal of Open Information Technologies 9.11 (2021): 68-74.

[8] Ilyushin, Eugene, Dmitry Namiot, and Ivan Chizhov. "Attacks on machine learning systems-common problems and methods." International Journal of Open Information Technologies 10.3 (2022): 17-22.

[9] Namiot, Dmitry. "Introduction to Data Poison Attacks on Machine Learning Models." International Journal of Open Information Technologies 11.3 (2023): 58-68.

[10] Намиот, Д. Е., Е. А. Ильюшин, and О. Г. Пилипенко. "Доверенные платформы искусственного интеллекта." International Journal of Open Information Technologies 10.7 (2022): 119-127.

[11] Song, Junzhe, and Dmitry Namiot. "On Real-Time Model Inversion Attacks Detection." International Conference on Distributed Computer and Communication Networks. Cham: Springer Nature Switzerland, 2023.

[12] Ribeiro, Mauro, Katarina Grolinger, and Miriam AM Capretz. "Mlaas: Machine learning as a service." 2015 IEEE 14th international conference on machine learning and applications (ICMLA). IEEE, 2015.

[13] Sedar, Roshan, et al. "A comprehensive survey of v2x cybersecurity mechanisms and future research paths." IEEE Open Journal of the Communications Society 4 (2023): 325-391.

[14] M. Uzair, "Who is liable when a driverless car crashes?", World Electric Veh. J., vol. 12, no. 2, 2021, [online] Available: https://www.mdpi.com/2032-6653/12/2/62.

[15] P. Penmetsa, P. Sheinidashtegol, A. Musaev, E. K. Adanu and M. Hudnall, "Effects of the autonomous vehicle crashes on public perception of the technology", IATSS Res., vol. 45, no. 4, pp. 485-492, 2021, [online] Available: https://www.sciencedirect.com/science/article/pii/S0386111221000224.

[16] M. Girdhar, J. Hong and J. Moore, "Cybersecurity of Autonomous Vehicles: A Systematic Literature Review of Adversarial Attacks and Defense Models," in IEEE Open Journal of Vehicular Technology, vol. 4, pp. 417-437, 2023, doi: 10.1109/OJVT.2023.3265363.

[17] Tesla Autopilot feature was involved in 13 fatal crashes, US regulator says https://www.theguardian.com/technology/2024/apr/26/tesla-autopilot-fatal-crash Retrieved: Jun, 2024

[18] G. Costantino and I. Matteucci, "Reversing Kia motors head unit to discover and exploit software vulnerabilities", J. Comput. Virol. Hacking Techn., vol. 19, pp. 33-49, 2022.

[19] Costantino, Gianpiero, Marco De Vincenzi, and Ilaria Matteucci. "A vehicle firmware security vulnerability: an IVI exploitation." Journal of Computer Virology and Hacking Techniques (2024): 1-16.

[20] Elkhail, Abdulrahman Abu, et al. "Vehicle security: A survey of security issues and vulnerabilities, malware attacks and defenses." IEEE Access 9 (2021): 162401-162437.
[21] Pham, Minh, and Kaiqi Xiong. "A survey on security attacks and defense techniques for connected and autonomous vehicles." Computers & Security 109 (2021): 102269.
[22] Ren, Huali, Teng Huang, and Hongyang Yan. "Adversarial examples: attacks and defenses in the physical world." International Journal of Machine Learning and Cybernetics 12.11 (2021): 3325-3336.
[23] Kurakin, Alexey, Ian J. Goodfellow, and Samy Bengio. "Adversarial examples in the physical world." Artificial intelligence safety and security. Chapman and Hall/CRC, 2018. 99-112.
[24] CIRCUMVENT FACIAL RECOGNITION WITH YARN https://hackaday.com/2023/04/16/circumvent-facial-recognition-with-yarn/ Retrieved: Jun, 2024
[25] Nassi, Ben, et al. "Phantom of the adas: Phantom attacks on driver assistance systems." Cryptology ePrint Archive (2020)
[26] «Мне в кабину стучат — ты нормальная?» Водители трамваев жалуются на автопилот, им отвечают статистикой https://www.fontanka.ru/2024/03/14/73330787/ Retrieved: Jun, 2024
[27] Hamdi, Mustafa Maad, et al. "A review on various security attacks in vehicular ad hoc networks." Bulletin of Electrical Engineering and Informatics 10.5 (2021): 2627-2635.
[28] Zhou, Husheng, et al. "Deepbillboard: Systematic physical-world testing of autonomous driving systems." Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering. 2020.
[29] Udacity Challenge. 2016. Steering angle model: Cg32. https://github.com/udacity/self-driving-car/tree/master/steeringmodels/community-models/cg23 (2016).
[30] Udacity Challenge. 2016. Steering angle model: Rambo. https://github.com/udacity/self-driving-car/tree/master/steeringmodels/community-models/rambo (2016).
[31] DeepBillboard https://github.com/deepbillboard/DeepBillboard Retrieved: Jun, 2024
[32] Patel, Naman, et al. "Adaptive adversarial videos on roadside billboards: Dynamically modifying trajectories of autonomous vehicles." 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2019.
[33] 9 Key Benefits of Mobile Billboard Advertising https://www.billups.com/articles/benefits-of-mobile-billboard-advertising Retrieved: Jun, 2024
[34] Sato, Takami, et al. "Security of deep learning based lane keeping system under physical-world adversarial attack." arXiv preprint arXiv:2003.01782 (2020).
[35] “OpenPilot: Open Source Driving Agent,” https://github.com/commaai/ openpilot, 2018
[36] https://sites.google.com/view/lane-keeping-adv-attack/ Retrieved: Jun, 2024
[37] Woitschek, Fabian, and Georg Schneider. "Physical adversarial attacks on deep neural networks for traffic sign recognition: A feasibility study." 2021 IEEE Intelligent vehicles symposium (IV). IEEE, 2021.
[38] Sitawarin, Chawin, et al. "Rogue signs: Deceiving traffic sign recognition with malicious ads and logos." arXiv preprint arXiv:1801.02780 (2018).
[39] Sitawarin, Chawin, et al. "Darts: Deceiving autonomous cars with toxic signs." arXiv preprint arXiv:1802.06430 (2018).
[40] Лентикулярная печать https://ru.wikipedia.org/wiki/%D0%9B%D0%B5%D0%BD%D1%82%D0%B8%D0%BA%D1%83%D0%BB%D1%8F%D1%80%D0%BD%D0%B0%D1%8F_%D0%BF%D0%B5%D1%87%D0%B0%D1%82%D1%8C Retrieved: Jun, 2024
[41] Morgulis, Nir, et al. "Fooling a real car with adversarial traffic signs." arXiv preprint arXiv:1907.00374 (2019).
[42] Han, Xingshuo, et al. "Physical backdoor attacks to lane detection systems in autonomous driving." Proceedings of the 30th ACM International Conference on Multimedia. 2022.
[43] Chernikova, Alesia, et al. "Are self-driving cars secure? evasion attacks against deep neural networks for steering angle prediction." 2019 IEEE Security and Privacy Workshops (SPW). IEEE, 2019.
[44] Arroyo, Miguel A., et al. "YOLO: frequently resetting cyber-physical systems for security." Autonomous Systems: Sensors, Processing, and Security for Vehicles and Infrastructure 2019. Vol. 11009. SPIE, 2019.
[45] Nguyen, Kien, et al. "Physical Adversarial Attacks for Surveillance: A Survey." IEEE Transactions on Neural Networks and Learning Systems (2023).
[46] Prishletsov, Dmitry, Sergey Prishletsov, and Dmitry Namiot. "Camouflage as adversarial attacks on machine learning models." International Journal of Open Information Technologies 11.9 (2023): 41-49.
[47] Zhang, Yang, et al. "CAMOU: Learning physical vehicle camouflages to adversarially attack detectors in the wild." International Conference on Learning Representations. 2018.
[48] Song, Dawn, et al. "Physical adversarial examples for object detectors." 12th USENIX workshop on offensive technologies (WOOT 18). 2018.
[49] Evtimov, Ivan, et al. "Robust physical-world attacks on machine learning models." arXiv preprint arXiv:1707.08945 2.3 (2017): 4.
[50] Kumar, K. Naveen, et al. "Black-box adversarial attacks in autonomous vehicle technology." 2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR). IEEE, 2020.
[51] Cao, Yulong, et al. "Adversarial sensor attack on lidar-based perception in autonomous driving." Proceedings of the 2019 ACM SIGSAC conference on computer and communications security. 2019.
[52] Namiot, D., and E. Ilyushin. "On Certification of Artificial Intelligence Systems." Physics of Particles and Nuclei 55.3 (2024): 343-346.
[53] Магистерская программа «Искусственный интеллект в кибербезопасности» (ФГОС) https://cs.msu.ru/node/3732 Retrieved: Jun, 2024
[54] Сухомлин, Владимир Александрович, et al. "Куррикулум дисциплины" Кибербезопасность"." (2022): 402-402.
[55] Сухомлин, Владимир Александрович. "Концепция и основные характеристики магистерской программы" Кибербезопасность" факультета ВМК МГУ." International Journal of Open Information Technologies 11.7 (2023): 143-148.
[56] Розничная торговля в цифровой экономике / В. П. Куприяновский, С. А. Синягов, Д. Е. Намиот [и др.] // International Journal of Open Information Technologies. – 2016. – Т. 4, № 7. – С. 1-12. – EDN WCMIWN.
[57] Развитие транспортно-логистических отраслей Европейского Союза: открытый BIM, Интернет Вещей и кибер-физические системы / В. П. Куприяновский, В. В. Аленьков, А. В. Степаненко [и др.] // International Journal of Open Information Technologies. – 2018. – Т. 6, № 2. – С. 54-100. – EDN YNIRFG.
[58] Искусственный интеллект как стратегический инструмент экономического развития страны и совершенствования ее государственного управления. Часть 2. Перспективы применения искусственного интеллекта в России для государственного управления / И. А. Соколов, В. И. Дрожжинов, А. Н. Райков [и др.] // International Journal of Open Information Technologies. – 2017. – Т. 5, № 9. – С. 76-101. – EDN ZEQDMT.

On Adversarial Attacks for Autonomous Vehicles

Dmitry Namiot, Vasily Kupriyanovsky, Alexey Pichugov

***Abstract*— This article examines adversarial attacks against machine (deep) learning models used in autonomous vehicles. Artificial intelligence (machine learning) systems play a decisive role in the functioning of unmanned vehicles. At the same time, all machine learning systems are susceptible to so-called adversarial attacks, when an attacker deliberately modifies data in such a way as to deceive the algorithms of such systems, complicate their work (reduce the quality of work), or achieve the behavior desired by the attacker. Adversarial attacks are a big problem for machine learning systems, especially when used in critical areas such as automated driving. Adversarial attacks pose a problem for functional testing - there is data on which the system does not work correctly (does not work at all, works with low quality). For autonomous vehicle systems, such attacks can be carried out in the physical form, when real objects captured by the vehicle's sensors are modified, dummy objects are created, etc. This article provides an overview of adversarial attacks on autonomous vehicles, focusing specifically on physical attacks.**

***Keywords*— machine learning, deep learning, adversarial attacks.**

REFERENCES

[1] NIST AI 100-2 E2023 Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. https://csrc.nist.gov/pubs/ai/100/2/e2023/final Retrieved: May, 2024

[2] Shibli, Ashfak Md, Mir Mehedi A. Pritom, and Maanak Gupta. "AbuseGPT: Abuse of Generative AI ChatBots to Create Smishing Campaigns." arXiv preprint arXiv:2402.09728 (2024).

[3] Namiot, Dmitry. "Schemes of attacks on machine learning models." International Journal of Open Information Technologies 11.5 (2023): 68-86.

[4] Bidzhiev, T. M., and D. E. Namiot. "Attacks on Machine Learning Models Based on the PyTorch Framework." Avtomatika i telemehanika 3 (2024): 38-50.

[5] Namiot, Dmitry, and Eugene Ilyushin. "Trusted Artificial Intelligence Platforms: Certification and Audit." International Journal of Open Information Technologies 12.1 (2024): 43-60.

[6] Namiot, Dmitry, Eugene Ilyushin, and Ivan Chizhov. "Ongoing academic and industrial projects dedicated to robust machine learning." International Journal of Open Information Technologies 9.10 (2021): 35-46. (in Russian)

[7] Namiot, D. E., E. A. Il'jushin, and I. V. Chizhov. "Osnovanija dlja rabot po ustojchivomu mashinnomu obucheniju." International Journal of Open Information Technologies 9.11 (2021): 68-74.

[8] Ilyushin, Eugene, Dmitry Namiot, and Ivan Chizhov. "Attacks on machine learning systems-common problems and methods." International Journal of Open Information Technologies 10.3 (2022): 17-22.

[9] Namiot, Dmitry. "Introduction to Data Poison Attacks on Machine Learning Models." International Journal of Open Information Technologies 11.3 (2023): 58-68.

[10] Namiot, D. E., E. A. Il'jushin, and O. G. Pilipenko. "Doverennye platformy iskusstvennogo intellekta." International Journal of Open Information Technologies 10.7 (2022): 119-127.

[11] Song, Junzhe, and Dmitry Namiot. "On Real-Time Model Inversion Attacks Detection." International Conference on Distributed Computer and Communication Networks. Cham: Springer Nature Switzerland, 2023.

[12] Ribeiro, Mauro, Katarina Grolinger, and Miriam AM Capretz. "Mlaas: Machine learning as a service." 2015 IEEE 14th international conference on machine learning and applications (ICMLA). IEEE, 2015.

[13] Sedar, Roshan, et al. "A comprehensive survey of v2x cybersecurity mechanisms and future research paths." IEEE Open Journal of the Communications Society 4 (2023): 325-391.

[14] M. Uzair, "Who is liable when a driverless car crashes?", World Electric Veh. J., vol. 12, no. 2, 2021, [online] Available: https://www.mdpi.com/2032-6653/12/2/62.

[15] P. Penmetsa, P. Sheinidashtegol, A. Musaev, E. K. Adanu and M. Hudnall, "Effects of the autonomous vehicle crashes on public perception of the technology", IATSS Res., vol. 45, no. 4, pp. 485-492, 2021, [online] Available: https://www.sciencedirect.com/science/article/pii/S0386111221000224.

[16] M. Girdhar, J. Hong and J. Moore, "Cybersecurity of Autonomous Vehicles: A Systematic Literature Review of Adversarial Attacks and Defense Models," in IEEE Open Journal of Vehicular Technology, vol. 4, pp. 417-437, 2023, doi: 10.1109/OJVT.2023.3265363.

[17] Tesla Autopilot feature was involved in 13 fatal crashes, US regulator says https://www.theguardian.com/technology/2024/apr/26/tesla-autopilot-fatal-crash Retrieved: Jun, 2024

[18] G. Costantino and I. Matteucci, "Reversing Kia motors head unit to discover and exploit software vulnerabilities", J. Comput. Virol. Hacking Techn., vol. 19, pp. 33-49, 2022.

[19] Costantino, Gianpiero, Marco De Vincenzi, and Ilaria Matteucci. "A vehicle firmware security vulnerability: an IVI exploitation." Journal of Computer Virology and Hacking Techniques (2024): 1-16.

[20] Elkhail, Abdulrahman Abu, et al. "Vehicle security: A survey of security issues and vulnerabilities, malware attacks and defenses." IEEE Access 9 (2021): 162401-162437.

[21] Pham, Minh, and Kaiqi Xiong. "A survey on security attacks and defense techniques for connected and autonomous vehicles." Computers & Security 109 (2021): 102269.

[22] Ren, Huali, Teng Huang, and Hongyang Yan. "Adversarial examples: attacks and defenses in the physical world." International Journal of Machine Learning and Cybernetics 12.11 (2021): 3325-3336.

[23] Kurakin, Alexey, Ian J. Goodfellow, and Samy Bengio. "Adversarial examples in the physical world." Artificial intelligence safety and security. Chapman and Hall/CRC, 2018. 99-112.

[24] CIRCUMVENT FACIAL RECOGNITION WITH YARN https://hackaday.com/2023/04/16/circumvent-facial-recognition-with-yarn/ Retrieved: Jun, 2024

[25] Nassi, Ben, et al. "Phantom of the adas: Phantom attacks on driver assistance systems." Cryptology ePrint Archive (2020)

[26] «Mne v kabinu stuchat — ty normal'naja?» Voditeli tramvaev zhalujutsja na avtopilot, im otvechajut statistikoj https://www.fontanka.ru/2024/03/14/73330787/ Retrieved: Jun, 2024

[27] Hamdi, Mustafa Maad, et al. "A review on various security attacks in vehicular ad hoc networks." Bulletin of Electrical Engineering and Informatics 10.5 (2021): 2627-2635.

[28] Zhou, Husheng, et al. "Deepbillboard: Systematic physical-world testing of autonomous driving systems." Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering. 2020.

[29] Udacity Challenge. 2016. Steering angle model: Cg32. https://github.com/udacity/self-driving-car/tree/master/steeringmodels/community-models/cg23 (2016).

[30] Udacity Challenge. 2016. Steering angle model: Rambo. https://github.com/udacity/self-driving-car/tree/master/steeringmodels/community-models/rambo (2016).

[31] DeepBillboard https://github.com/deepbillboard/DeepBillboard Retrieved: Jun, 2024

[32] Patel, Naman, et al. "Adaptive adversarial videos on roadside billboards: Dynamically modifying trajectories of autonomous vehicles."

2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2019.
[33] 9 Key Benefits of Mobile Billboard Advertising https://www.billups.com/articles/benefits-of-mobile-billboard-advertising Retrieved: Jun, 2024
[34] Sato, Takami, et al. "Security of deep learning based lane keeping system under physical-world adversarial attack." arXiv preprint arXiv:2003.01782 (2020).
[35] “OpenPilot: Open Source Driving Agent,” https://github.com/commaai/ openpilot, 2018
[36] https://sites.google.com/view/lane-keeping-adv-attack/ Retrieved: Jun, 2024
[37] Woitschek, Fabian, and Georg Schneider. "Physical adversarial attacks on deep neural networks for traffic sign recognition: A feasibility study." 2021 IEEE Intelligent vehicles symposium (IV). IEEE, 2021.
[38] Sitawarin, Chawin, et al. "Rogue signs: Deceiving traffic sign recognition with malicious ads and logos." arXiv preprint arXiv:1801.02780 (2018).
[39] Sitawarin, Chawin, et al. "Darts: Deceiving autonomous cars with toxic signs." arXiv preprint arXiv:1802.06430 (2018).
[40] Lentikuljarnaja pechat' https://ru.wikipedia.org/wiki/%D0%9B%D0%B5%D0%BD%D1%82%D0%B8%D0%BA%D1%83%D0%BB%D1%8F%D1%80%D0%BD%D0%B0%D1%8F_%D0%BF%D0%B5%D1%87%D0%B0%D1%82%D1%8C Retrieved: Jun, 2024
[41] Morgulis, Nir, et al. "Fooling a real car with adversarial traffic signs." arXiv preprint arXiv:1907.00374 (2019).
[42] Han, Xingshuo, et al. "Physical backdoor attacks to lane detection systems in autonomous driving." Proceedings of the 30th ACM International Conference on Multimedia. 2022.
[43] Chernikova, Alesia, et al. "Are self-driving cars secure? evasion attacks against deep neural networks for steering angle prediction." 2019 IEEE Security and Privacy Workshops (SPW). IEEE, 2019.
[44] Arroyo, Miguel A., et al. "YOLO: frequently resetting cyber-physical systems for security." Autonomous Systems: Sensors, Processing, and Security for Vehicles and Infrastructure 2019. Vol. 11009. SPIE, 2019.
[45] Nguyen, Kien, et al. "Physical Adversarial Attacks for Surveillance: A Survey." IEEE Transactions on Neural Networks and Learning Systems (2023).
[46] Prishletsov, Dmitry, Sergey Prishletsov, and Dmitry Namiot. "Camouflage as adversarial attacks on machine learning models." International Journal of Open Information Technologies 11.9 (2023): 41-49.
[47] Zhang, Yang, et al. "CAMOU: Learning physical vehicle camouflages to adversarially attack detectors in the wild." International Conference on Learning Representations. 2018.
[48] Song, Dawn, et al. "Physical adversarial examples for object detectors." 12th USENIX workshop on offensive technologies (WOOT 18). 2018.
[49] Evtimov, Ivan, et al. "Robust physical-world attacks on machine learning models." arXiv preprint arXiv:1707.08945 2.3 (2017): 4.
[50] Kumar, K. Naveen, et al. "Black-box adversarial attacks in autonomous vehicle technology." 2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR). IEEE, 2020.
[51] Cao, Yulong, et al. "Adversarial sensor attack on lidar-based perception in autonomous driving." Proceedings of the 2019 ACM SIGSAC conference on computer and communications security. 2019.
[52] Namiot, D., and E. Ilyushin. "On Certification of Artificial Intelligence Systems." Physics of Particles and Nuclei 55.3 (2024): 343-346.
[53] Magisterskaja programma «Iskusstvennyj intellekt v kiberbezopasnosti» (FGOS) https://cs.msu.ru/node/3732 Retrieved: Jun, 2024
[54] Suhomlin, Vladimir Aleksandrovich, et al. "Kurrikulum discipliny" Kiberbezopasnost'"." (2022): 402-402.
[55] Suhomlin, Vladimir Aleksandrovich. "Koncepcija i osnovnye harakteristiki magisterskoj programmy" Kiberbezopasnost'" fakul'teta VMK MGU." International Journal of Open Information Technologies 11.7 (2023): 143-148.
[56] Roznichnaja torgovlja v cifrovoj jekonomike / V. P. Kuprijanovskij, S. A. Sinjagov, D. E. Namiot [i dr.] // International Journal of Open Information Technologies. – 2016. – T. 4, # 7. – S. 1-12. – EDN WCMIWN.
[57] Razvitie transportno-logisticheskih otraslej Evropejskogo Sojuza: otkrytyj BIM, Internet Veshhej i kiber-fizicheskie sistemy / V. P. Kuprijanovskij, V. V. Alen'kov, A. V. Stepanenko [i dr.] // International Journal of Open Information Technologies. – 2018. – T. 6, # 2. – S. 54-100. – EDN YNIRFG.
[58] Iskusstvennyj intellekt kak strategicheskij instrument jekonomicheskogo razvitija strany i sovershenstvovanija ee gosudarstvennogo upravlenija. Chast' 2. Perspektivy primenenija iskusstvennogo intellekta v Rossii dlja gosudarstvennogo upravlenija / I. A. Sokolov, V. I. Drozhzhinov, A. N. Rajkov [i dr.] // International Journal of Open Information Technologies. – 2017. – T. 5, # 9. – S. 76-101. – EDN ZEQDMT.