ISSN 2225-9309 (Print)
2024. № 2 (31). С. 40–43

Молодежный
Вестник УГАТУ

http://journal.ugatu.su

УДК 004.93

СОСТЯЗАТЕЛЬНЫЕ АТАКИ НА СИСТЕМЫ РАСПОЗНАВАНИЯ ОБРАЗОВ

Е. А. Григорьев[1], А. Д. Минаева[2], И. С. Ефимов[3]

[1] egor.ufa.rb@mail.ru, [2] am20016@bk.ru, [3] efimov903@mail.ru

[1-3] ФГБОУ ВО «Уфимский университет науки и технологий» (УУНиТ)

Аннотация: в данной статье представлен обзор состязательных атак и методов защиты от них. В данной статье обсуждаются подходы к состязательным атакам. Состязательные атаки могут быть нацелены на критически важные системы, такие как специализированные приложения, беспилотные транспортные средства и другие. Эти нападения являются наиболее опасными и представляют большую угрозу.

Ключевые слова: состязательная атака; машинное обучение; информационная безопасность; уязвимости; методы защиты.

ВВЕДЕНИЕ

За последние 10 лет искусственный интеллект активно развивался и привлекал внимание различных стран и IT-гигантов, а также наращивал исследования и инвестиции в эту сферу. Машинное обучение пока добилось больших успехов не только в области компьютерного зрения, распознавания речи и обработки естественного языка, но и в области беспилотных автомобилей и биометрической аутентификации. К сожалению, этот прогресс также привел к росту интернет-преступности, связанной с использованием искусственного интеллекта.

Атаки на системы с искусственным интеллектом могут иметь серьезные последствия, например, когда системы компьютерного зрения ошибочно идентифицируют объекты на улице или в системах безопасности. Это требует серьезного внимания и исследований в области защиты от подобных атак, так как угроза атак противника на системы компьютерного зрения реальна.

ОПРЕДЕЛЕНИЕ ПОНЯТИЯ СОСТЯЗАТЕЛЬНОЙ АТАКИ

Состязательная атака (Adversarial attacks), вредоносное машинное обучение (Adversarial Machine Learning, AML) — это концепция, которая подразумевает атаку на системы машинного обучения с целью изменения входных данных таким образом, чтобы нейронные сети неправильно их понимали [1-2].

Благодаря широкому использованию машинного обучения в 2013 году ученые сделали интересное открытие - нейронные сети можно легко обмануть малейшими изменениями в данных – возмущениями. Классическая иллюстрация состязательной атаки представлена на рис. 1.

ПРИМЕРЫ СОСТЯЗАТЕЛЬНЫХ АТАК

В настоящее время модели компьютерного зрения достигли уровня, сопоставимого или даже превосходящего человеческое зрение. Они используются в самых разных областях, таких как автономные транспортные средства, распознавание лиц, медицинская диагностика, системы видеонаблюдения и обнаружение вредоносных программ.

До сих пор исследователи обучали и тестировали модели машинного обучения в контролируемых средах, в рамках конкурсов и научных исследованиях. Однако при реализации в реальном сценарии ошибки модели могут привести к проблемам безопасности.

Известно, что даже небольшие изменения в изображениях могут разрушить системы машинного обучения и неправильно классифицировать изображения. Эти специально созданные изображения, известные как «состязательные примеры», являются одним из известных ограничений глубокого обучения.

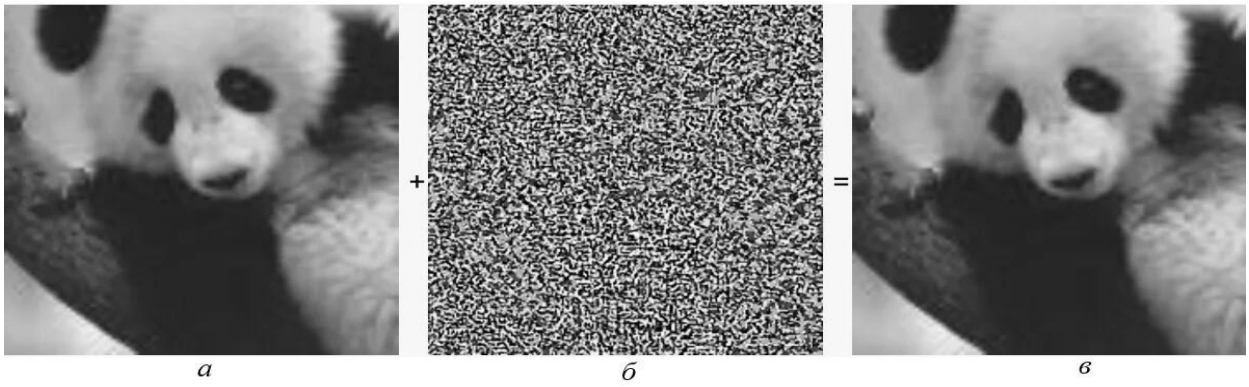

Рис. 1. Обман нейросети возмущением:
а – исходное изображение, классифицированное как панда с вероятностью 57,7%
б - небольшое возмущение, незаметное для человека, но замечаемое нейросетью
в – незаметно измененное изображение, классифицированное как гиббон с вероятностью 99,3%

“Сила” глубокого обучения заключается в способности распознавать закономерности в данных. Если показать нейросети десятки тысяч помеченных фотографий животных, она научится определять, какие паттерны(образцы) связаны с пандами, а какие - с обезьянами. Затем она может применить эти знания для выявления новых изображений животных, которых она никогда не встречала.

Но модели глубокого обучения также очень хрупки. Системы распознавания изображений полагаются только на пиксельные рисунки, а не на более глубокое понимание того, что они видят, поэтому легко сломать эти шаблоны и заставить их видеть что-то совершенно другое.

Алгоритмы выявления рака кожи могут ошибочно связывать наличие маркеров на снимке с проявлениями меланомы [3]. Это связано с тем, что линейка присутствует на всех изображениях со злокачественными поражениями на коже, и алгоритму легче научиться распознавать линейку как образец, а не различать различные типы пятен на коже.

Иногда такие паттерны могут быть более тонкими и ненавязчивыми. Например, когда каждая камера имеет свой цифровой след.

КЛАССИФИКАЦИЯ И ВИДЫ СОСТЯЗАТЕЛЬНЫХ АТАК

В современной научной литературе выделяют несколько основных категорий атак, направленных на создание состязательных атак [4]:

- создание состязательного возмущения включает в себя добавление небольших помех к ранее правильно классифицированному изображению, что приводит к неправильной классификации изображения. Такое вмешательство, изменяющее классификацию модели, называется состязательным;

- использование состязательных патчей. Выбранная область изображения полностью заменяется специально выбранной заплаткой(патчем). Обычно предполагается, что патч закрывает сам классифицируемый объект, а содержание, представленное на фото, все же понятно человеку. Но для нейронных сетей это становилось проблемой, и они могли начать неправильно классифицировать объект;

- использование абстрактных состязательных примеров - бессмысленные для человека образы (например, похожие на белый шум или некие абстрактные паттерны), но нейросети уверенно классифицируют образы на определенные классы. Такое присваивание нелогичных классов считается нежелательным и необычным поведением;

- физическая состязательная атака (также известная как атака экземпляра) - изменение объектов реального мира, которые после фотографирования, неверно классифицируются нейронной сетью.

В зависимости от способа изменения объекта атаки можно выделить три типа атак:

- состязательные патчи — это небольшие изображения, которое можно прикрепить к поверхности атакуемого объекта. Этот метод прост в реализации, но наиболее эффективен на плоских поверхностях;
- камуфляжные атаки предназначены для маскировки 3D-объектов с помощью специально разработанных текстур, которые можно наносить на поверхности, например, в качестве рисунка на одежде или машине;
- в случае оптической атаки злоумышленник может использовать осветительное оборудование, такое как проектор или лазерный излучатель, чтобы направить свет или лазер на цель и, таким образом, изменить ее внешний вид, точно контролируя и маскируя атаку. Однако они чувствительны к освещению окружающей среды, что влияет на их использование в реальности.

Существует простой способ разделить состязательные атаки на две основные категории: целевые и нецелевые атаки [5]:

- если модель присваивает состязательному примеру класс, отличный от правильного, нецелевая атака считается успешной;
- целевые атаки считаются успешными только в том случае, если модель классифицирует состязательные примеры на заранее выбранные классы ошибок.

Другим важным элементом классификации является «знания злоумышленника»:

- если у злоумышленника есть подробная информация о внутренней структуре модели - «вредоносные знания»: как данные готовятся к обучению, где они находятся, какие функции выполняет атакуемая система, какой алгоритм используется и каков результат и пр., значит атака происходит в режиме белого ящика;
- если атакующий не обладает подробными знаниями архитектуры и параметров модели, атака происходит в режиме чёрного ящика. При этом предполагается возможность взаимодействовать с моделью и наблюдать ее реакцию на ввод. Данный вид атаки считается наиболее распространенным.

МЕТОДЫ ЗАЩИТЫ ОТ СОСТЯЗАТЕЛЬНЫХ АТАК

В свете появления угрозы состязательных атак, требуется пересмотр стратегий обучения нейронных сетей.

Однако повышение качества обучающих данных не является полным решением для устранения риска состязательных атак. Для оценки уровня безопасности алгоритмов машинного обучения требуется многомерный подход, включающий следующие этапы [6]:

- выявление потенциальных слабых мест в алгоритмах машинного обучения в процессе обучения и классификации;
- выбор соответствующих мер противодействия выявленным угрозам и оценка воздействия на целевые систем;
- разработка стратегии противодействия потенциальным атакам.
- традиционные методы повышения стабильности моделей машинного обучения (такие как снижение веса и фильтрация) обычно не обеспечивают реальной защиты от враждебных действий. Поэтому эмпирический метод защиты актуален, а его эффективность проверена и подтверждена на практике;
- защитная дистилляция: это метод обучения модели прогнозированию вероятности различных классов, вместо того чтобы принимать сложные решения о том, к какому классу она принадлежит. Вероятность обеспечивается предыдущей моделью, обученной с использованием той же задачи, но с использованием сложного класса. Это позволяет создать модель с более гладкой поверхностью, затрудняя злоумышленнику обнаружение враждебных данных, которые приводят к неправильной классификации;
- дополнительные классы означают, что классификатор обучается на основе определенного распределения данных в пределах определенного предела, что позволяет избежать идентификации неизвестных классов.

Таким образом, состязательные атаки показывают, что многие модели машинного обучения могут быть разрушены различными способами. В конце концов, состязательное машинное обучение создает

множество сложнейших проблем для искусственного интеллекта и машинного обучения, и это относительно новая область сетевой безопасности. Универсального решения для защиты этих моделей от различных атак не существует, но в будущем могут появиться более совершенные методы и разумные стратегии для решения этой серьезной задачи.

ЗАКЛЮЧЕНИЕ

В заключение следует отметить, что состязательные атаки на системы распознавания образов представляют серьёзную угрозу для безопасности и надёжности этих систем. Они могут привести к неправильному распознаванию объектов, а также к утечкам данных и другим негативным последствиям.

Для защиты от таких атак необходимо применять меры по повышению устойчивости систем распознавания образов к состязательным воздействиям. Это может включать в себя использование более сложных моделей и алгоритмов, обучение на более разнообразных наборах данных, а также применение методов обнаружения и предотвращения атак.

Дальнейшие исследования в области состязательных атак и методов защиты от них являются важными для обеспечения безопасности систем распознавания образов и других технологий искусственного интеллекта.

СПИСОК ЛИТЕРАТУРЫ

1. Жариков И.А., Черняк М.В. Адверсариальные атаки на нейронные сети // Интеллектуальные системы. - 2019. - № 2(28). - С. 15-18.

2. «Как обмануть нейросеть или что такое Adversarial attack» // 2023 [Электронный ресурс]. URL: https://chernobrovov.ru/articles/kak-obmanut-nejroset-ili-chto-takoe-adversarial-attack.html / (Дата обращения: 12.12.2023).

3. Бурцев М.И., Гупалин Н.В., Ярцев И.А. Adversarial attacks on machine learning models // Учебное пособие. - 2020. - Издательство: Издательский дом НИУ ВШЭ. - 218 с.

4. Герасимов А.С., Воронов И.А. Адверсариальные атаки на нейронные сети в задачах компьютерного зрения // Молодежный научный форум, Интернет-журнал. - 2019. - Т. 15, № 3. - С. 72-78.

5. Савицкий А.С., Гомзин К.А. Адверсариальные атаки на нейронные сети: обзор и исследование // Сборник трудов Нижегородского государственного университета им. Н.И. Лобачевского. - 2019. - Т. 3(39), № 1. - С. 67-73.

6. Гордиенко Ю.А., Казаковцева Н.С. Адверсариальные атаки на нейронные сети: особенности и методы защиты // Актуальные проблемы информатики и вычислительной техники. - 2019. - Т. 21, № 3. - С. 302-307.

ОБ АВТОРАХ

ГРИГОРЬЕВ Егор Андреевич, магистрант каф.ВТиЗИ. Дипл. бакалавр (УУНиТ, 2023).

МИНАЕВА Анастасия Дмитриевна, магистрант каф.ВТиЗИ. Дипл. бакалавр (УУНиТ, 2023).

ЕФИМОВ Иван Сергеевич, магистрант каф.ВТиЗИ. Дипл. бакалавр (УУНиТ, 2023).

METADATA

Title: Adversarial attacks on image recognition systems

Authors: E. A. Grigoriev[1], A. D. Minaeva[2], I. S. Efimov[3]

Affiliation:

[1-3] Ufa State Aviation Technical University (UGATU), Russia.

Email: [1] egor.ufa.rb@mail.ru, [2] am20016@bk.ru, [3] efimov903@mail.ru

Language: Russian.

Source: Molodezhnyj Vestnik UGATU (scientific journal of Ufa University of Science and Technology), no. 2 (31), pp. 40-43, 2024. ISSN 2225-9309 (Print).

Abstract: This article provides an overview of adversarial attacks and methods of protection against them. This article discusses approaches to adversarial attacks. Adversarial attacks can target mission-critical systems such as specialized applications, unmanned vehicles, and others. These attacks are the most dangerous and pose the greatest threat.

Key words: adversarial attack; machine learning; information security; vulnerabilities; protection methods.

About authors:

GRIGORIEV Egor Andreevich, master's student of the department of VTiZI. Bachelor's Degree (UUST, 2023).

MINAEVA Anastasia Dmitrievna, master's student of the department of VTiZI. Bachelor's Degree (UUST, 2023).

EFIMOV Ivan Sergeevich, master's student of the department of VTiZI. Bachelor's Degree (UUST, 2023).