АТАКИ И МЕТОДЫ ЗАЩИТЫ В СИСТЕМАХ МАШИННОГО ОБУЧЕНИЯ: АНАЛИЗ СОВРЕМЕННЫХ ИССЛЕДОВАНИЙ

Котенко И. В.[1], Саенко И. Б.[2], Лаута О. С.[3], Васильев Н. А.[4], Садовников В. Е.[5]

DOI: 10.21681/2311-2024-1-24-37

Цель исследования: *проведение анализа атак на системы машинного обучения и методов защиты от них на основе известных обзорных статей, опубликованных за последние пять лет в высокорейтинговых журналах.*

Методы исследования: *системный анализ, классификация, моделирование, машинное обучение.*

Полученные результаты: *исследованы обзорные работы в высокорейтинговых журналах, посвященные анализу атак на системы машинного обучения и методам защиты от них. Выявлено, что тематика защиты от таких атак вызывает в настоящее время постоянно растущий интерес, а сфера воздействия таких атак охватывает интеллектуальные системы различного предназначения с ориентацией на широкий спектр типов обрабатываемых данных (изображения, звук, текст, видео, кибербезопасность и т.д.). Обобщены признаки классификации атак на системы машинного обучения и мер защиты от них. Выделены и рассмотрены наиболее распространенные атаки, которые по своему типу относятся к атакам «белого ящика» или «черного ящика». Обоснованы наиболее распространенные методы защиты от атак на системы машинного обучения, и дана их характеристика. Для ряда наиболее сложных методов защиты приведено их детальное описание на уровне отдельных этапов. Выделены особенности реализации методов защиты, позволяющие повысить эффективность обнаружения атак на системы машинного обучения.*

Научная новизна: *анализ работ по тематике атак на системы машинного обучения и мер защиты от них показал, что в настоящее время для них не существует устоявшейся классификации, что обусловлено бурным ростом новых разновидностей атак и появлением новых методов и механизмов защиты. Предложенные в рассмотренном исследовании признаки классификации атак и методов защиты обобщают подходы к такой классификации. Описание наиболее распространенных методов защиты отличается от других известных описаний поэтапной детализацией, которая обеспечивает простоту реализации этих методов в системах защиты интеллектуальных системах различного назначения.*

Вклад: *Котенко И. В. и Саенко И. Б. – общая концепция анализа атак на системы машинного обучения и методов защиты от них на основе известных обзорных работ; Котенко И. В. и Лаута О. С. – классификация и характеристика атак; Васильев Н. А. и Садовников В. Е. – классификация и поэтапная детализация мер защиты; Котенко И. В. и Саенко И. Б. – обсуждение особенностей реализации методов защиты.*

Ключевые слова: *кибербезопасность, машинное обучение, глубокое обучение, состязательные атаки, защита от атак, искусственный интеллект.*

ATTACKS AND DEFENSE METHODS IN MACHINE LEARNING SYSTEMS: ANALYSIS OF MODERN RESEARCH

Igor Kotenko[6], Igor Saenko[7], Oleg Lauta[8], Nikita Vasiliev[9], Vladimir Sadivnikov[10]

The purpose of the study: *conducting an analysis of attacks on machine learning systems and methods of protection against them based on well-known review works published in recent years in highly rated journals.*

Research methods: *system analysis, classification, modeling, machine learning.*

1 Котенко Игорь Витальевич, заслуженный деятель науки РФ, доктор технических наук, профессор, главный научный сотрудник и руководитель лаборатории проблем компьютерной безопасности, ФГБУН «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук» (СПб ФИЦ РАН), г. Санкт-Петербург, Россия. E-mail: ivkote@comsec.spb.ru

2 Саенко Игорь Борисович, доктор технических наук, профессор, ведущий научный сотрудник, ФГБУН «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук» (СПб ФИЦ РАН), г. Санкт-Петербург, Россия. E-mail: ibsaen@comsec.spb.ru

3 Лаута Олег Сергеевич, доктор технических наук, доцент, Государственный университет морского и речного флота им. адмирала С. О. Макарова (ГУМРФ), г. Санкт-Петербург, Россия. E-mail: laos-82@yandex.ru

4 Васильев Никита Алексеевич, научный сотрудник, Военная академия связи им. Маршала Советского Союза С.М. Будённого, г. Санкт-Петербург, Россия. E-mail: vasn2020@mail.ru

5 Садовников Владимир Евгеньевич, аспирант, ФГБУН «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук» (СПб ФИЦ РАН), г. Санкт-Петербург, Россия. E-mail: bladimir1998@mail.ru

6 Igor Kotenko, Honored Worker of Science of the Russian Federation, Doctor of Technical Sciences, Professor, Chief Scientist and Head of Laboratory of Computer Security Problems at St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg, Russia. E-mail: ivkote@comsec.spb.ru

7 Igor Saenko, Doctor of Technical Sciences, Professor, Leading researcher of Laboratory of Computer Security Problems at St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg, Russia. E-mail: ibsaen@comsec.spb.ru

8 Oleg Lauta, Doctor of Technical Sciences, associate professor, Admiral Makarov State University of Maritime and Inland Shipping, St. Petersburg, Russia. E-mail: laos-82@yandex.ru

9 Nikita Vasiliev, researcher, Military Telecommunication Academy named after the Soviet Union Marshal Budienny S.M., St. Petersburg, Russia. E-mail: vasn2020@mail.ru

10 Vladimir Sadovnikov, graduate student of Laboratory of Computer Security Problems at St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg, Russia. E-mail: bladimir1998@mail.ru

Results obtained: *review papers in high-ranking journals devoted to the analysis of attacks against machine learning systems and methods of protection against them were studied. It was revealed that the topic of protection against such attacks is currently of growing interest, and the scope of influence of such attacks covers intelligent systems for various purposes, focusing on a wide range of types of processed data (images, sound, text, video, etc.). The characteristics of the classification of attacks on machine learning systems and measures to protect against them were summarized. The most common attacks, which in their type belong to «white box» or «black box» attacks, were identified and considered. The most common methods of protection against attacks against machine learning systems are substantiated and their characteristics were given. For the most complex protection methods, their detailed description was given at the level of individual stages. Features of the implementation of protection methods were highlighted that make it possible to increase the efficiency of detecting attacks against machine learning systems.*

Scientific novelty: *an analysis of works on the topic of attacks against machine learning systems and measures to protect against them has shown that currently there is no established classification for them, which is due to the rapid growth of new types of attacks and the emergence of new protection methods and mechanisms. The characteristics of the classification of attacks and defense methods presented in the work generalize the approaches to such classification proposed in the works studied. The description of the most common protection methods differs from other known descriptions in its step-by-step detailing, which ensures ease of implementation of these methods in protection systems for intelligent systems for various purposes.*

Contribution: *Igor Kotenko and Igor Saenko – general concept of analysis of attacks against machine learning systems and methods of protection against them based on well-known review works; Igor Kotenko and Oleg Lauta – classification and characteristics of attacks; Nikita Vasilev and Vladimir Sadovnikov – classification and step-by-step detailing of protection measures; Igor Kotenko and Igor Saenko – discussion of the features of implementing protection methods.*

Keywords: *cyber security, machine learning, deep learning, adversarial attacks, attack protection, artificial intelligence.*

Введение

Искусственный интеллект (ИИ) и машинное обучение (МО) являются областями компьютерной науки, которые все шире используются для создания систем поддержки принятия решений, способных обрабатывать и анализировать информацию таким же образом, как это делает человек. Ключевой идеей функционирования систем МО является то, что они способны самостоятельно обучаться, анализируя большие объемы опытных данных, а затем использовать полученные знания для принятия решений в новых ситуациях. Эффективность систем МО обусловлена их способностью намного быстрее, чем человек, находить скрытые закономерности в данных, описывающих различные процессы или явления. Благодаря этому системы МО получают все более широкое внедрение во многих сферах деятельности человека [1].

В медицине системы МО используются для обработки медицинских данных, диагностики заболеваний и разработки индивидуальных планов лечения [1, 2]. В финансовой сфере системы МО применяются для прогнозирования рыночных трендов, оптимизации инвестиционных стратегий и улучшения управления рисками [4]. В промышленности системы МО применяются для автоматизации процессов, оптимизации планирования и управления качеством продукции [5]. На транспорте системы МО используются для оптимизации маршрутов и планирования логистики, улучшения безопасности на дорогах и разработки автономных транспортных средств [6]. В энергетике использование методов МО позволяет оптимизировать производство и планирование ресурсов, прогнозировать возможные перегрузки в электрической сети и принимать меры по более эффективному управлению энергией [7]. В сфере образования системы МО применяются для разработки адаптивных образовательных платформ, индивидуализации обучения, оптимизации пути обучения, выработки рекомендаций для улучшения результатов обучения [8].

Атаки на системы МО, иначе называемые состязательными (adversarial), могут приводить к нарушению безопасности и надежности функционирования целевых систем. Большинство состязательных атак сводится к изменению обучающего набора данных МО или изменению самой модели МО. Это может вызвать искажение результатов, полученных с применением технологии МО, что в критических системах приведет к серьезным последствиям и большим экономическим убыткам. Поэтому защита систем МО от атак становится в настоящее время одной из важнейших задач в области информационной безопасности и находит отражение в большом количестве научных трудов, посвященных этой тематике.

Количество работ, в которых рассматриваются состязательные атаки и методы защиты от них, возрастает с каждым годом. При этом каждый год появляются работы, в которых уточняются алгоритмы реализации таких атак и рассматриваются новые решения по противодействию этим атакам. Авторами настоящей работы также предлагались различные подходы и решения по защите от атак на системы МО [9–11].

В то же время представляет несомненный интерес проведение глубокого анализа известных атак и методов защиты в системах машинного обучения. Для этой цели авторы проанализировали обзорные работы по состязательным атакам и методам защиты от них, появившиеся в последние годы. Всего было отобрано для анализа свыше 30 обзорных работ, опубликованных за последние пять лет. Рассмотрение этих работ позволило выделить основные тренды в совершенствовании состязательных атак и методов защиты.

Целью настоящей работы является анализ атак на системы МО и методы защиты от них, проведенный на основе рассмотрения обзорных статей, посвященных состязательным атакам. Новизна работы заключается в представлении основных результатов и трендов в области реализации атак на системы МО и защиты от таких атак в рамках одной работы. В статье дается характеристика наиболее используемых состязательных атак и методов защиты. Рассмотрение методов защиты детализировано до алгоритмического уровня. Это позволяет данной статье помочь исследователям в области информационной безопасности систем МО.

Анализ работ по тематике атак на системы МО и методов защиты от них

Отбор работ для анализа осуществлялся следующим образом. Были задействованы поисковые системы Google.ru и Scholar.google.com (Академия Google). Использовалась строка поиска «Adversarial attack defense review». Период поиска был установлен в диапазоне 2020–2023 годы. Отбору подлежали обзорные статьи, опубликованные в журналах, имеющих квартиль Q1 – Q3 в международной индексной базе Scopus. Квартиль определялся по данным, представленным в системе https://www.scimagojr.com. Всего таким образом было отобрано для анализа 24 обзорные статьи. В таблице 1 приведены общие сведения об отобранных статьях.

Таблица 1

Общие сведения об отобранных обзорных статьях

П1	П2	П3	П4	П5	П6
[12]	2020	Q1	203	11	Глубокое обучение
[13]	2020	Q1	72	18	Вредоносный код
[14]	2021	Q2	139	25	Машинное обучение
[15]	2020	Q2	136	12	Изображения, графика, текст
[16]	2020	Q3	115	4	Глубокое обучение
[17]	2020	Q3	13	4	Электроэнергия
[18]	2021	Q1	450	339	Компьютерное зрение
[19]	2021	Q1	152	34	Умные грид-сети
[20]	2021	Q1	78	1	Глубокое обучение
[21]	2021	Q2	163	38	Кибербезопасность
[22]	2021	Q2	65	42	Умные грид-сети
[23]	2021	Q3	132	75	Изображения, текст, вредоносный код
[24]	2022	Q1	185	121	Глубокое обучение
[25]	2022	Q1	128	45	Глубокое обучение
[26]	2022	Q1	52	21	Глубокое обучение
[27]	2022	Q1	46	26	Цифровые сигналы
[28]	2022	Q1	34	27	Глубокое обучение
[29]	2022	Q2	103	48	Глубокое обучение
[30]	2022	Q2	49	21	Умные грид-сети
[31]	2022	Q2	46	7	Изображения
[32]	2023	Q1	246	140	Автономный транспорт
[33]	2023	Q1	176	119	Текст
[34]	2023	Q1	53	23	Кибербезопасность
[35]	2023	Q2	179	48	Графика

В таблице 1 учитываются следующие показатели: П1 – ссылка; П2 – год публикации; П3 – квартиль по базе Scopus; П4 – общее количество ссылок в статье; П5 – количество ссылок на источники, опубликованные с 2019 года; П6 – предметная область.

Анализируя данные по показателю П3, можно сделать вывод, что выбранные работы по тематике атак на системы МО имеют большую значимость и вызывают несомненный интерес. Подавляющее большинство работ было опубликовано в высокорейтинговых журналах квартилей Q1 и Q2.

Показатели П4 и П5 характеризуют широту охвата выбранных работ. В журналах квартилей Q1 и Q2, как правило, показатель П4 принимает значение, большее чем 100 (хотя есть несколько работ со значением гораздо ниже). Показатель П5, по нашему мнению, характеризует актуальность обзоров. Для работ, опубликованных в 2020 году, показатель П5 имеет небольшое значение, не превышающее 20, что является вполне естественным. Для следующих годов этот показатель увеличивается, принимая наивысшие значения для 2023 года. Это говорит о том, что, как правило, в различных обзорных работах можно найти ссылки на одни и те же исходные публикации, в которых рассматриваются конкретные случаи реализации различных типов состязательных атак и методов защиты.

Анализ показателя П6 позволяет сделать вывод, что большинство обзоров имеют широкую предметную область, которая отмечена как «глубокое обучение» или «машинное обучение» (9 обзоров). В то же время имеются обзоры, посвященные конкретной предметной области. В качестве таких предметных областей рассматриваются: изображения, графика, текст и/или вредоносный код – 6 работ; электроэнергия и умные грид-сети – 4; кибербезопасность – 2; компьютерное зрение – 1; цифровые сигналы – 1; автономный транспорт – 1. Такой широкий разброс предметных областей в обзорных статьях говорит о том, что тематика состязательных атак и защиты от них затрагивает в настоящее время практически все современные интеллектуальные системы.

В каждой из обзорных работ рассматривались классификация атак и мер защиты, а также давались характеристика и примеры их реализации, взятые из оригинальных источников. При этом следует отметить, что классификация атак и мер защиты из года в год уточнялась.

Далее рассмотрим наиболее распространенные атаки на системы машинного обучения и методы защиты от них, выделенные в результате анализа отобранных статей.

Атаки на системы МО

Для классификации атак на системы МО в работах [12–35] используются различные признаки, в частности, следующие:

- ✓ метод атаки (атаки на части модели МО или прямые атаки на данные);
- ✓ область данных, на которые направлена атака (изображения, звук, текст и т.д.);
- ✓ цель атаки (компрометация модели, уклонение от детектирования и т.д.);
- ✓ тип входных данные (непрерывные или дискретные);
- ✓ владение знаниями об атакуемой системе («белый ящик», «черный ящик», «серый ящик») и другие признаки.

Последний признак использовался в системах классификации атак, предлагаемых практически во всех работах. Поэтому в настоящей работе остановимся на рассмотрении этого признака.

Значение «белый ящик» означает, что злоумышленник полностью владеет необходимыми знаниями и о модели МО, и об обучающих наборах данных. «Черный ящик» – противоположный случай, когда у атакующего нет информации ни о модели МО, ни о наборах данных. Промежуточным вариантом является «серый ящик», когда злоумышленник владеет частичными, неполными знаниями о модели МО и наборах данных.

В таблице 2 представлено распределение наиболее известных состязательных атак по значениям признака владения знаниями. Название атаки дается в оригинале на английском языке, затем следует перевод и принятое для этой атаки сокращение.

Рассмотрим более подробно наиболее популярные методы атаки на системы МО.

FGSM (быстрый метод, основанный на знаке градиента) – это метод атаки «белого ящика» на нейронные сети, который используется для обмана моделей, обученных для распознавания изображений. Метод FGSM заключается в том, чтобы изменить изображение незначительно таким образом, чтобы обученная модель ошибочно идентифицировала его другим классом. Для этого используется метод градиентного спуска, который позволяет найти наиболее чувствительные пиксели на изображении. При использовании метода FGSM начальное изображение рассматривается как точка на пути от исходного до измененного изображения, которое обеспечивает максимальное изменение значения скорости потерь целевой функции (функции потерь). Затем производится вычисление градиента потерь по отношению к каждому пикселю изображения, после чего все пиксели с наименьшим модулем градиента

Таблица 2

Распределение атак по значениям признака владения знаниями

Значение признака	Название атаки (англ.)	Перевод названия	Сокращение
Белый ящик	Fast Gradient Sign Method	Метод быстрого градиента	FGSM
	Iterative Gradient Sign Method	Итеративный градиентный метод	IGSM
	Jacobian Saliency Map Attack	Атака карты значимости на основе Якобиана	JSMF
	Basic Iterative Method	Базовый итеративный метод	BIM
	Undetectable Perturbation	Незначительные изменения	UP
	Carlini and Wagner's Attack	Атака Карлини и Вагнера	C&W
	Iterative Least-Likely Class Method	Итеративный метод класса с наименьшей вероятностью	ILCM
	One-Step Target Class Method	Метод одношагового целевого класса	OSTCM
	Deep Fool	«Полный дурак»	DF
	Hot/Cold method	Горячий/холодный метод	HCM
	Ground-Truth Attack	Истина о системе	GTA
Черный ящик	Boundary Attack	Граничная атака	BA
	Zero-Query Attacks	Атака с нулевым запросом	ZQA
	Generative Adversarial Network	Генеративно-состязательная сеть	GAN
	One Pixel Attack	Атака одним пикселем	OPA
	Zeroth Order Optimization	Оптимизация нулевого порядка	ZOO
	Genetic Algorithms	Генетические алгоритмы	GA
	Improved Genetic Algorithm	Улучшенный генетический алгоритм	IFA
	Probability Weighted Word Saliency	Вероятностно-взвешенная значимость слова	PWWS
	Greedy Search Algorithm	Жадный алгоритм поиска	GSA
	Natural Evolution Strategies	Естественные эволюционные стратегии	NES
	Insertion and Removal of Words	Вставка и удаление слов	IRW
	Real-World Noise	Шум реального мира	RWN
Серый ящик	Cross-Site Scripting	Межсайтовый скриптинг	CSS
	Password Guessing	Подбор паролей	PG
	Buffer Overflow Attack	Атака переполнения буфера	BOA
	SQL Injection	SQL-инъекция	SQLI
	Weak Authentication Attack	Атака слабой аутентификации	WAA
	Cross-Site Request Forgery	Межсайтовая подделка запроса	CSRF

обнуляются, а остальные увеличиваются или уменьшаются на значение, которое составляет знак градиента. Метод FGSM позволяет создавать поддельные изображения, которые выглядят практически также, как оригиналы, но несут с собой измененную информацию, которая может обмануть модель машинного обучения. Существует несколько вариантов FGSM, которые отличаются тем, как определяется величина шага градиентного спуска. Например, FGSM может использоваться с фиксированным шагом либо определять шаг в каждой точке с применением линейного поиска с обратным ходом (backtracking line search). У FGSM есть ограничения. Одно из них заключается в том, что метод может обмануть модель только до определенной степени, после чего результаты перестают быть достоверными, и модель начинает идентифицировать измененное изображение правильно. Кроме того, FGSM может быть применен только к моделям, которые используют градиентный спуск для обучения. Следует также отметить, что методы атаки, такие как FGSM, могут быть использованы не только злоумышленниками, но и для различных исследовательских задач, связанных с оценкой уровня защиты нейронных сетей и их поведения в различных сценариях. В частности, метод FGSM может быть использован для разработки новых алгоритмов защиты нейронных сетей, позволяющих повышать уровень защиты от подобных атак.

IGSM (итеративный метод, основанный на знаке градиента) является разновидностью метода атаки «белого ящика» на нейронные сети, который расширяет возможности схожего алгоритма FGSM. Он основан на многократном применении метода FGSM с учетом нескольких изменений. IGSM является алгоритмом оптимизации, который начинается с исходного изображения и продолжает обновлять его через серию итераций с использованием FGSM. В каждой итерации значения пикселей изменяются в направлении увеличения потерь целевой функции. В отличие от FGSM, который использует только одну итерацию для создания поддельных изображений, IGSM повторяет процедуру атаки на каждой итерации, что дает лучший эффект, но требует больших вычислительных ресурсов. IGSM может быть использован как для целевой, так и для нецелевой атаки.

JSMA (атака карты значимости на основе Якобиана) – это алгоритм атаки «белого ящика» на системы определения поддельных изображений, основанный на методах глубокого обучения. Этот алгоритм использует вектор градиентов (Якобиан), который определяет, как изменения весов в нейронной сети повлияют на выходной результат. В результате JSMA может определить наиболее «важные» признаки (части) изображения, которые влияют на классификацию модели. JSMA начинается с выбора целевой модели для атаки. Затем вычисляется вектор градиентов для каждой части изображения, позволяющий определить те части, которые влияют на классификацию картинки как подделку. Затем увеличивается влияние этих частей изображения, уменьшая влияние других частей, и приводя к результату, когда нейронная сеть классифицирует подделку.

BIM (базовый итеративный метод) – это тип атаки «белого ящика» на системы МО, который основан на внедрении изменений во входные данные. Результаты атаки могут привести к ошибочным выводам или неправильным действиям системы. Примеры BIM-атак включают изменение значений входных параметров при обучении моделей машинного обучения. Например, если система обучается определять различные образцы на основе цвета, размера и формы, то злоумышленник может ее обмануть, предоставив входные данные, содержащие измененные значения цвета, размера и формы. Другой пример BIM-атаки может быть направлен на автоматизированные системы контроля качества, когда злоумышленник отправляет измененные данные, создавая ложные сигналы об ошибке. Такие атаки могут вызвать сбои в системе, неправильную работу оборудования или опасные сбои в производственном процессе. Для предотвращения BIM-атак необходимо включать меры безопасности при разработке и настройке систем МО, такие как проверка входных данных, использование контроля целостности данных и обучение моделей на большом количестве данных. Также следует использовать методы дополнительного контроля, такие как применение одноразовых ПИН-кодов или двухфакторной аутентификации.

UP-атака (введение незначительных изменений) – это тип атаки «белого ящика», который заключается во внедрении незначительных изменений в данные или параметры модели, приводящих к ошибочным выводам и дискредитации результатов. Основная цель такой атаки – обойти систему защиты и создать искаженные данные, чтобы они были приняты за правильные. UP-атака может быть использована в различных областях, например, для манипулирования результатами голосования, изменения прогнозов погоды, машинного обучения, автономных транспортных средств, медицинских диагностических систем. Она является достаточно сложной для обнаружения, так как создает незначительные изменения в данных.

BA (граничная атака) – это типовой метод атаки «черного ящика», основанный на принятии решений. Начиная с исходного состязательного изображения, в нем используется бинарный поиск для нахождения точки выборки, которая находится вблизи границы классификации. Производится случайное блуждание по границе между двумя противоположными областями, чем уменьшается расстояние от целевого изображения. В соответствии с этим шагом продолжается итерация и постепенно уменьшается расстояние от исходного изображения. Причина, по которой этот тип алгоритма называется «граничной атакой», заключается в том, что он генерирует состязательные примеры путем поиска вдоль границы до тех пор, пока они не сойдутся для получения оптимального или рационального решения. Результаты, полученные таким методом, могут удовлетворять требованиям ошибочной классификации модели «черного ящика». Общее возмущение, которое увеличивается по сравнению с исходным изображением, зависит от производительности алгоритма.

ZQA (атака с нулевым запросом) – это атака «черного ящика», которая предназначена для передачи опыта между моделями без доступа к информации входных данных. В их основе лежит передача знаний между моделями, используя выводы моделей, а не входные данные. Традиционно для передачи опыта между моделями требуется доступ к исходным данным моделей, что может привести к утечке конфиденциальной информации. Злоумышленники могут применять атаки ZQA для выполнения различных задач, например, для создания фальшивых изображений и видео, которые позволяют обмануть системы компьютерного зрения, или для атак на защищенные системы распознавания лица, используя

данные полученные от других систем распознавания. Кроме того, атаки ZQA могут быть использованы для идентификации конфиденциальной информации. Например, злоумышленники могут использовать их для обнаружения ключевых слов и фраз в документах, которые не должны быть доступны публично. Они могут использовать знания, полученные от одной модели, чтобы обучить другую модель, которая может идентифицировать эти конфиденциальные данные. В то же время, атаки ZQA могут быть использованы и в благих целях. Например, они могут использоваться для передачи опыта между моделями в области медицины или научных исследований, позволяя ускорить процесс обучения модели и позволить получить более точные результаты. Защита от ZQA может быть построена на основе использования методов обнаружения аномалий и обучения с учителем. Эти методы могут идентифицировать необычные выходные данные, которые могут быть связаны с ZQA. Другим возможным направлением защиты является разработка методов обнаружения и предотвращения передачи опыта между моделями, используя только выходные данные. Защиту можно усилить с помощью обучения моделей для препятствования реализации атак и замедлять процесс передачи опыта между моделями.

Атака с использованием GAN (генеративно-состязательной сети) – это метод атаки «черного ящика», использующий нейронные сети GAN для формирования различных атак на модели МО. Принцип работы GAN заключается в тренировке двух нейронных сетей – генератора и дискриминатора, которые последовательно передают друг к другу данные и обучаются. На первом этапе генератор создает поддельные примеры данных, которые передаются дискриминатору вместе с настоящими примерами из обучающего набора. Дискриминатор обучается отличать настоящие данные от поддельных, и генератор учится создавать такие данные, чтобы их было сложно отличить от реальных. На втором этапе генератор использует полученные знания о структуре данных, чтобы создать атаки на модель машинного обучения. Эти атаки могут быть различными в зависимости от типа модели и задачи, которую она решает. Как только атака сгенерирована, она может быть использована злоумышленником для нападения на модель МО. Таким образом, GAN позволяет генерировать различные виды атак на модели МО, что делает их более уязвимыми для нападений. Это может быть использовано для тестирования устойчивости моделей и нахождения уязвимостей в их защите.

OPA (атака одним пикселем) относится к атакам «черного ящика» и основывается на алгоритмах МО. Она использует уязвимости в работе нейронных сетей, которые определяют изображения на основе цветовых значений каждого пикселя. Основной принцип работы этой атаки состоит в том, чтобы изменить значение всего лишь одного пикселя на изображении таким образом, чтобы нейронная сеть неправильно классифицировала это изображение. Например, при редактировании фото с котом, атака OPA может изменить значение пикселя на месте носа кота таким образом, что нейросеть будет считать, что на самом деле изображается собака. Атака OPA использует эволюционные алгоритмы, позволяющие определить оптимальные пиксели и изменить их значения таким образом, чтобы обмануть нейронную сеть. Использование таких алгоритмов позволяет достичь максимальной эффективности атаки при минимальном числе изменений на изображении.

NES (естественные эволюционные стратегии) – это семейство алгоритмов численной оптимизации для задач «черного ящика». Как и все другие эволюционные стратегии, они итеративно обновляют параметры поискового распределения, следуя естественному градиенту в сторону более высокой ожидаемой приспособленности. Общая процедура заключается в следующем. Для создания множества точек поиска используется параметризованное распределение. В каждой точке оценивается функция соответствия. Параметры распределения позволяют алгоритму адаптивно фиксировать значения функции приспособленности. Например, в случае распределения Гаусса они включают среднее значение и ковариационную матрицу. На основе выборок оценивается градиент поиска в сторону более высокой ожидаемой пригодности. Затем выполняется шаг подъема вдоль естественного градиента. Этот шаг имеет решающее значение, так как он предотвращает колебания, преждевременное схождение и нежелательные эффекты, возникающие из-за заданной параметризации. Весь процесс повторяется до тех пор, пока не будет выполнен критерий останова.

Методы защиты от атак на системы МО

Возможными признаками классификации методов защиты от атак на МО, указанными в работах [12–35], являются:

- ✓ направленность защиты (наборы данных, модель МО);
- ✓ способ анализа наборов данных (обнаружение изменений, защита от предобработки);
- ✓ способ обработки модели МО (обнаружение состязательных примеров, укрепление модели);
- ✓ направленность на слой нейронной сети (входной слой, промежуточные слои, выходной слои) и другие.

В результате можно выделить следующие наиболее популярные методы защиты (рис. 1):

1) состязательная тренировка (competitive training);
2) оборонительная дистилляция (defensive distillation);
3) реконструкция входных данных (input data reconstruction);
4) фреймворк Defense-GAN;
5) подкрепление (укрепление) модели (model reinforcement);
6) защита от предварительной обработки (protection from preprocessing);
7) обнаружение примеров состязательности (detection of adversarial examples).

Среди перечисленных методов только Defense-GAN реализуется с использованием соответствующего фреймворка. Остальные методы могут быть реализованы на основе применения различных средств.

Рассмотрим подробнее содержание этих методов защиты.

Рис. 1. Основные методы защиты от атак на системы МО

Состязательная тренировка. «Состязательная тренировка» – это метод защиты от атак и взломов систем МО, который использует GAN. Как было сказано выше, нейронная сеть GAN состоит из двух компонентов: генератора и дискриминатора. Генератор создает поддельные данные, а дискриминатор учится отличать их от настоящих. Генератор учится создавать данные таким образом, чтобы они были максимально похожи на настоящие. Дискриминатор учится отличать их от настоящих с большой точностью. Этот процесс продолжается до тех пор, пока генератор не научится создавать данные, которые дискриминатор не сможет отличить от настоящих. Таким образом, система МО, использующая состязательную тренировку, может обучиться отличать подделки от настоящих данных, что делает ее более защищенной от атак и взломов.

В процессе реализации метода состязательной тренировки можно выделить следующие этапы (рис. 2).

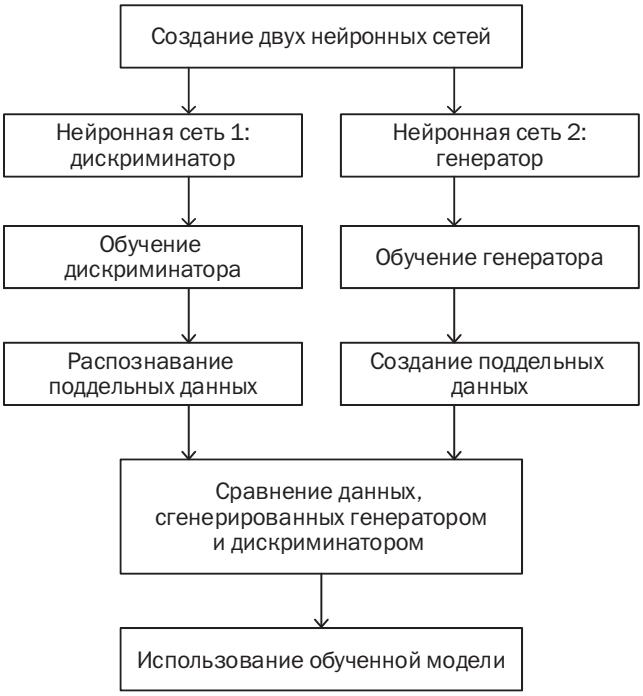

Рис. 2. Этапы реализации метода состязательной тренировки

Этап 1. Создание двух нейронных сетей: генератора и дискриминатора.

Этап 2. Обучение генератора созданию поддельных данных (изображений, звуков, текстов и т.д.). Параллельно дискриминатор обучается отличать поддельные данные от настоящих.

Этап 3. Генератор и дискриминатор конкурируют между собой в рамках задачи, которая была определена для тренировки модели (например, распознавание лиц).

Этап 4. Генератор создает новые поддельные данные, а дискриминатор оценивает, насколько они похожи на реальные данные. Оценка дискриминатора передается обратно генератору, чтобы тот мог улучшить свои навыки.

Этап 5. Дискриминатор обучается становиться все более точным в распознавании подделок. Генератор, в свою очередь, учится создавать все более качественные подделки. Этот процесс повторяется множество раз, пока подделки не станут практически неотличимыми от настоящих данных.

Этап 6. Модель, полученная после тренировки, может быть использована для анализа новых данных. Например, она может использоваться для распознавания лиц на фотографиях, прикрепленных к заявкам на кредит, для выявления лжи при беседе с клиентом и т.д.

В зависимости от приложений и областей, в которых применяется этот метод, шаги могут незначительно изменяться для достижения лучших результатов. Однако общий подход остается прежним и заключается в конкурентной тренировке генератора и дискриминатора с целью создания лучших поддельных данных и защиты систем МО от атак и взломов.

Оборонительная дистилляция. Оборонительная дистилляция – это метод защиты, базирующийся на создании и использовании так называемой «отфильтрованной» или «дистиллированной» версии данных, которую алгоритмы МО могут использовать для обучения своих моделей. Данные проходят через фильтры и алгоритмы, которые идентифицируют и удаляют такие элементы данных, которые могут быть опасными для системы. Например, это могут быть данные, содержащие вредоносный код или данные со специальными символами, которые предназначены для реализации атак на систему.

Оборонительная дистилляция, как правило, включает в себя несколько этапов: определение набора данных для обучения, анализ и фильтрацию данных, создание обучающих моделей и их эксплуатацию.

Набор данных, используемый для оборонительной дистилляции, должен быть репрезентативным и содержать данные, которые могут быть использованы для определения и обнаружения типов атак и уязвимостей в системе. Эти данные подвергаются анализу и фильтрации. После фильтрации данных они используются для обучения моделей МО, которые, в свою очередь, могут использоваться для обнаружения потенциальных угроз безопасности и принятия соответствующих мер.

Существенным преимуществом оборонительной дистилляции является ее способность к защите системы от новых видов атак. Традиционные методы защиты, такие как фильтрация трафика и использование антивирусов, направлены преимущественно на обнаружение и блокирование известных угроз, в то время как оборонительная дистилляция способна обнаружить и защитить систему от новых видов атак, которые еще не известны.

Реконструкция входных данных. Метод реконструкции входных данных основан на идее создания механизма защиты, который позволяет анализировать входные данные и осуществлять их реконструкцию, которая затем сравнивается с начальными входными данными. Если входные данные были изменены злоумышленником, то реконструкция будет отличаться от начальных данных, что позволит сигнализировать о возможности атаки на систему. Однако необходимо учитывать, что этот метод может иметь высокий уровень ложных срабатываний.

Процесс реализации метода реконструкции входных данных можно разделить на следующие этапы.

Этап 1. Обработка входных данных. На этом этапе входные данные проходят предварительную обработку, например, они могут быть преобразованы в числовой вид, аномалии и шум могут быть удалены.

Этап 2. Создание реконструкции на основании модели, используемой для обучения, работы системы МО и в соответствии с правилами обработки данных.

Этап 3. Сравнение реконструкции с изначальными входными данными. Если восстановленные данные отличаются от исходных данных, то предупреждение о том, что, возможно, система была атакована, входные данные были заменены или изменены. Другая причина – вероятность ошибки при выполнении алгоритма превышает разумный уровень.

Этап 4. Принятие соответствующих мер. Например, можно прекратить обучение или работу, попросить у пользователя подтверждение правильности входных данных или оповестить администратора о возможной атаке.

Важным аспектом работы метода реконструкции входных данных является выбор модели, используемой для создания реконструкции. Эта модель должна быть способна точно восстанавливать входные данные при минимальной потере информации. Выбор модели зависит от конкретной задачи и особенностей данных.

Defense-GAN. Defense-GAN – это фреймворк, предназначенный для защиты от атак на GAN-сети. Defense-GAN настраивает защищаемые модели путем генерации видоизмененных исходных данных, что делает атакующую модель недействительной, так как она обучается на искаженной информации.

Для определения эффективности искажений Defense-GAN использует статистические метрики, которые оценивают, насколько хорошо искажения защищают модель от атаки. Если метрики показывают, что защищаемая модель имеет хорошую защиту от атак, то генерируемые искажения можно использовать для защиты от нежелательных воздействий.

Кроме того, Defense-GAN использует критерии обучения, направленные на защиту, которые основаны на минимизации потерь при классификации и систематическом сдвиге наиболее важных признаков на изображении. Эти критерии обучения помогают защищаемой модели лучше предсказывать классы, а также улучшают ее устойчивость к атакам.

В процессе защиты с использованием фреймворка Defense-GAN можно выделить отдельные этапы (рис. 3).

Этап 1 (подготовка данных). Этот этап включает загрузку данных и их предварительную обработку. Для обучения фреймворка необходимы наборы

изображений, которые будут использоваться для обучения генератора и классификатора. Создание такого набора может включать в себя множество шагов предварительной обработки, таких как изменение размеров, повороты, зеркальные отражения и другие.

Этап 2 (обучение генератора). На этом этапе фреймворк обучается создавать защищенные версии исходных изображений на основе целевой функции, используя GAN-генератор. Генератор обучается создавать новые, но похожие на исходные изображения, которые будут менее чувствительны к различным типам атак.

Этап 3 (обучение дискриминатора). После обучения генератора начинается обучение дискриминатора – классификатора, который будет использоваться для оценки качества изображений. Классификатор обучается давать наиболее точную оценку классов, которым принадлежат изображения, а также определять, насколько защищены изображения.

Этап 4 (тестирование). После обучения модели проводится ее тестирование на тестовом наборе данных, который не использовался при обучении. Это позволяет оценить способность фреймворка защищать изображения от различных типов атак и определить, насколько точным является классификатор.

Этап 5 (защита от атак). При обнаружении атаки на модель глубокого обучения механизмы защиты могут включаться автоматически, используя защищенные версии изображений, созданные GAN-генератором. Это может помочь защитить модель от перебросок искажений, уменьшения качества изображений и других типов атак.

Этап 6 (оценка качества). Оценка качества модели осуществляется путем оценки ее способности защитить модель глубокого обучения от разных типов атак. При этом могут использоваться такие метрики, как достоверность (accuracy) и F-мера.

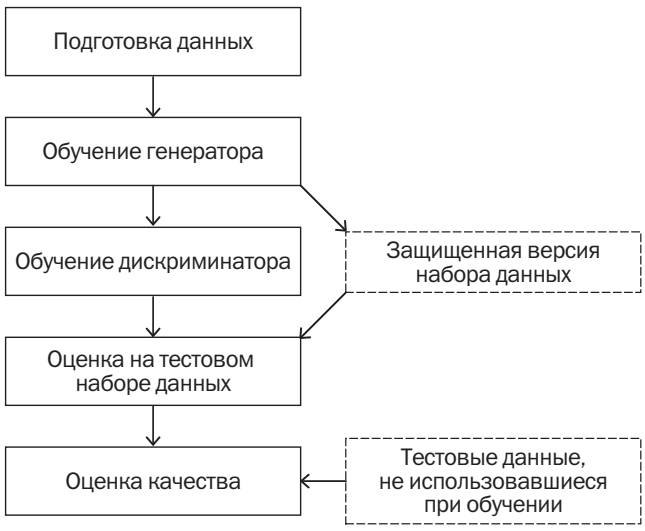

Рис. 3. Этапы работы фреймворка Defense-GAN

Подкрепление модели МО. Подкрепление модели МО – это метод защиты, который заключается в том, чтобы создать дополнительный слой защиты, добавляющий «укрепление» к системе, защищая ее от атак, ориентированных на модель. Идея заключается в использовании уже созданных моделей, которые определяют, какие действия являются безопасными, а какие – потенциально вредными. Система может использовать эти модели, чтобы определить, какие действия должны быть разрешены, и какие – запрещены. Если действие не соответствует модели безопасности, то оно будет заблокировано.

Один из примеров использования метода подкрепления модели – это алгоритм деревьев решений. Он использует деревья для принятия решений, основанных на определенных факторах безопасности. Если действия пользователя соответствуют модели, то алгоритм разрешит их. Если нет, то действия будут заблокированы.

Защита от предварительной обработки данных. Защита от предварительной обработки данных, также известная как «защита трансформаций», представляет собой метод защиты компьютерных систем от атак, основанных на предварительной обработке данных перед их отправкой на сервер. Метод предполагает изменение формы или контента сообщения перед их отправкой на сервер с целью затруднить анализ данных злоумышленником.

Этапы метода защиты от предварительной обработки данных (защиты трансформаций) представлены на рис. 4.

Этап 1 (анализ данных). Сначала компьютерная система анализирует данные, которые будут отправлены на сервер. Это могут быть, например, данные, введенные пользователем на веб-странице или в приложении.

Этап 2 (трансформация данных). Затем система трансформирует данные таким образом, чтобы изменить их форму или контент. Например, это может быть замена символов на другие символы или добавление случайных данных, чтобы затруднить распознавание оригинальных данных.

Этап 3 (отправка данных на сервер). После трансформации данные отправляются на сервер для дальнейшей обработки.

Этап 4 (распознавание трансформации). Когда данные приходят на сервер, система должна распознать трансформацию, которая была применена к данным.

Этап 5 (обработка данных). После распознавания трансформации система должна обработать данные с учетом трансформации. Это может включать в себя дешифрование данных или удаление случайной информации, которая была добавлена к данным.

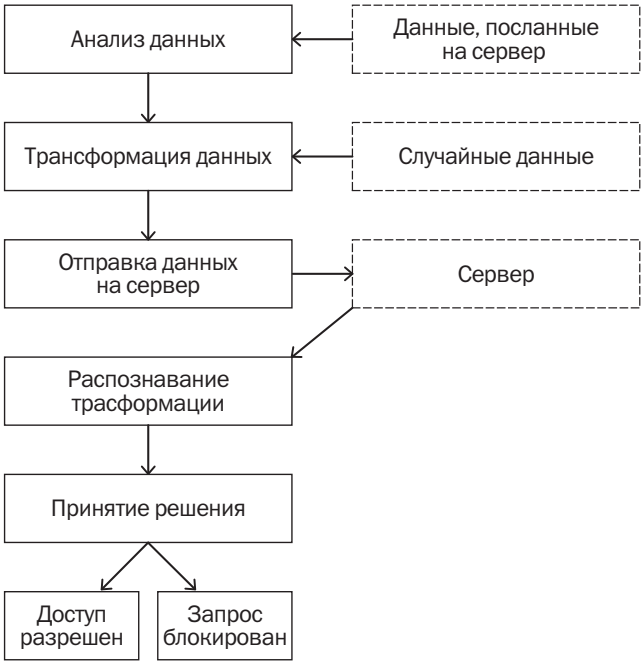

Рис. 4. Этапы метода защиты от предварительной обработки данных

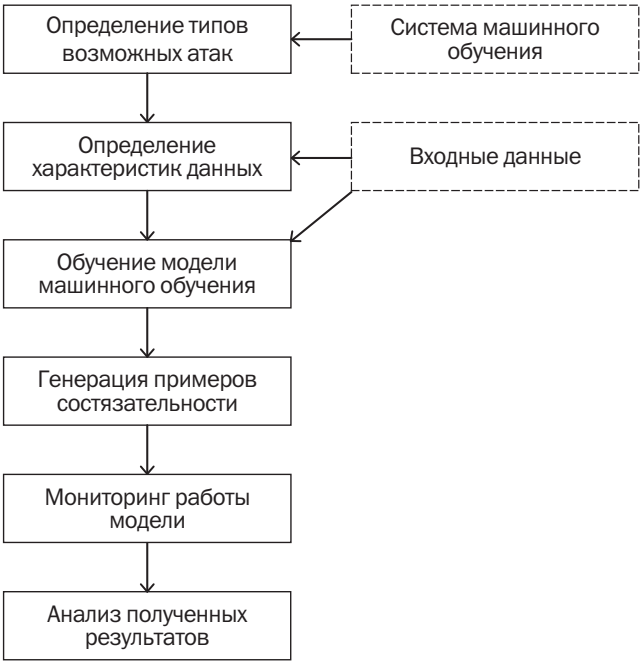

Рис.5. Этапы метода обнаружения примеров состязательности

Этап 6 (принятие решения). На основе обработанных данных система принимает решение, как обрабатывать запрос. Например, она может разрешить доступ к определенным ресурсам или заблокировать запрос, если он не соответствует политикам безопасности.

В целом, метод защиты от предварительной обработки данных может помочь защитить компьютерную систему от многих видов атак, которые основываются на предварительной обработке данных. Однако он не является единственным и должен использоваться в сочетании с другими методами.

Обнаружение примеров состязательности. Метод обнаружения примеров состязательности (Adversarial Examples) заключается в поиске и обнаружении таких примеров данных, которые могли бы использоваться для атак на модель МО. Примеры состязательности – это измененные данные, которые были подготовлены для того, чтобы обмануть модель МО. Например, это могут быть изображения, на которых были внесены незначительные изменения, которые затрудняют или делают невозможным их распознавание моделью.

Метод обнаружения примеров состязательности может быть реализован с использованием следующих подходов:

1) мониторинг необычных изменений в поведении модели МО; если модель начинает давать неправильные ответы или сильно изменяется ее точность, это может быть признаком того, что она подвергается атаке состязательных примеров;
2) анализ данных на предмет выявления изменений в распределении; примеры состязательности могут изменить распределение входных данных, что может быть замечено при анализе статистических показателей;
3) использование тестовых наборов данных с примерами состязательности для обучения модели; если модель обучена на данных с такими примерами, она может стать более устойчивой к таким атакам в будущем.

Метод может снижать свою точность при работе с обычными данными. Поэтому необходим баланс между защитой от примеров состязательности и сохранением высокой точности модели.

Процесс мониторинга необычных изменений в работе модели МО для обнаружения примеров состязательности делится на следующие этапы (рис. 5).

Этап 1 (определение типов возможных атак). Необходимо определить, какие типы примеров состязательности могут быть применены для атаки на модель МО, чтобы затем разработать методы обнаружения таких атак.

Этап 2 (определение характеристик данных). Следует провести анализ входных данных, на основе которых модель МО принимает решения, и выявить их наиболее значимые характеристики и признаки.

Этап 3 (обучение модели МО). Нужно обучить модель МО на основе данных, которые не содержат примеров состязательности, чтобы получить базовую версию модели.

Этап 4 (генерация примеров состязательности). Необходимо сгенерировать различные примеры

состязательности, которые могут быть использованы для атаки на модель МО, на основе знаний о характеристиках входных данных и типах возможных атак.

Этап 5 (мониторинг работы модели). Нужно наблюдать за изменениями в работе модели МО, которые могут свидетельствовать о наличии входных данных, содержащих примеры состязательности. Это может происходить на основе анализа метрик качества обучения, таких как точность, время обработки или показатели ошибок.

Этап 6 (анализ и документация результатов). Необходимо проанализировать полученные результаты и задокументировать эффективность методов обнаружения примеров состязательности, чтобы повысить эффективность системы защиты.

Проблема защиты систем машинного обучения от атак

Искусственный интеллект и, в частности, машинное/глубокое обучение являются мощным инструментом в сфере информационной безопасности. Эти технологии могут быть использованы как для защиты систем, так и для реализации атак. Применение МО может значительно улучшить методы защиты систем. МО и анализ данных позволяют разрабатывать более сложные и инновационные методы по обнаружению атак и предотвращению угроз.

Однако при использовании МО при реализации атак возникает серьезная проблема: системы защиты могут стать бессильными перед алгоритмами, использующими МО. Реализация атак на системы защиты, основанные на методах МО, представляет серьезную угрозу. Например, атакующая сторона может использовать нейронные сети для создания фальшивых данных и обмана системы защиты. Это приводит к неправильным решениям или проникновению в систему через механизмы защиты, которые не смогут идентифицировать подобные атаки.

Фреймворки, основанные на МО, способны выявлять аномальное поведение и реагировать на новые виды атак, которые ранее были неизвестны. Стоит заметить, что несмотря на все достоинства фреймворка Defense-GAN, у него существует следующий недостаток. Отсутствие зависимости от точки инициализации нейронной сети в прикладных задачах защиты информации влечет за собой то, что оптимальный дискриминатор будет присваивать более высокое значение для функции потерь, чем самим частям реальных обрабатываемых данных из генератора. Если модифицировать работу дискриминатора, то сам дискриминатор сразу же станет неоптимальным. По этой причине функционирование фреймворка Defense-GAN должно быть скомпоновано с соответствующей функцией потерь, которая будет учитывать описанную выше особенность для той предметной области знаний, в рамках которой решается задача обеспечения информационной безопасности.

Метод OPA весьма неэффективен, если в базовой конструкции модели машинного обучения используется два и более слоя пулинга. При этом неважно, какой тип пулинга используется, поскольку комбинация слоев данного типа нивелирует эффект разности пикселей на уровне высокоуровневых признаков.

Нужно признать, что метод оборонительной дистилляции имеет особенности при защите моделей ансамблей. Так, если в решающей модели будет использоваться алгоритм с привилегированной информацией, то раздельное функционирование модели-ученика и модели-учителя может привести к коллизиям в процессе нормальной работы базовой модели.

Метод JSMA также обладает следующей важной особенностью. Он не может функционировать одновременно с моделью МО. Поэтому правильная организация потоков данных в пайп-лайне построения модели МО поможет полностью исключить негативный эффект от внедрения JSMA в качестве вредоносного компонента. При этом не потребуется каких-либо дополнительных надстроек, контролирующих процесс функционирования основной модели МО.

Наконец, следует заметить, что защита от атак является весьма нетривиальной. Основная модель МО может функционировать неправильно при появлении в наборе данных аномальных величин. К аномальным показателям относятся не только попытки взлома, но и, чаще всего, просто некоторые редкие значения. Например, рост человека в 240 см может показаться аномальным для обобщающей способности алгоритма, если модель чаще всего оперировала с данными среднестатистического роста взрослого человека. По этой причине внедрять компоненты защиты от атак следует с особой осторожностью, так как нестандартные данные могут быть интерпретированы системой защиты как попытка несанкционированного доступа к данным.

Заключение

В статье представлены результаты анализа обзорных работ по реализации атак (состязательных атак) и методам защиты в системах машинного обучения, опубликованных за 2020–2023 годы в высокорейтинговых журналах. Анализ показал, что тематика защиты от состязательных атак вызывает в настоящее время постоянно растущий интерес и наблюдается бурный рост исследований в этой области. Причем эта сторона информационной безопасности касается различных видов обрабатываемых данных (изображения, звук, текст, компьютерное зрение и т.д.) и различных сфер жизни человека (медицина, транспорт, финансы, экономика и т.д.).

В ходе анализа атак на системы МО выделены наиболее часто используемые в обзорах признаки их классификации и дана общая характеристика наиболее распространенных атак, которые по своему типу относятся к атакам «белого ящика», «черного ящика» и «серого ящика». В качестве наиболее распространенных атак «белого ящика» выделены атаки типов FGSM, IGSM, JSMF, BIM и UP. В качестве наиболее распространенных атак «черного ящика» выделены атаки типов BA, ZQA, GAN и OPA.

Выделены признаки классификации методов защиты от состязательных атак и дана характеристика наиболее распространенных методов. К числу таких методов были отнесены состязательная тренировка, оборонительная дистилляция, реконструкция входных данных, фреймворк Defense-GAN, подкрепление модели, защита от предварительной обработки и обнаружение примеров состязательности. Для наиболее сложных методов защиты приведено их детальное описание на уровне отдельных этапов.

Результаты проведенного анализа могут быть использованы для разработки новых эффективных методов и механизмов защиты от угроз, связанных с технологией МО. Например, для создания методики, способной оценивать и выбирать методы защиты систем МО, что в настоящее время является одной из ключевых проблем в сфере информационной безопасности.

Рецензент: *Липатников Валерий Алексеевич, доктор технических наук, профессор, научный сотрудник научно-исследовательского центра Военной академии связи имени Маршала Советского Союза С. М. Буденного, Санкт-Петербург, Россия. E-mail: lipatnikovanl@mail.ru*

Литература

1. *Клименко Р. В., Тарароев Я. В. Философское осмысление применения технологий машинного обучения. Перспективы искусственного интеллекта // Социальное время. 2016. № 1 (5). С. 15–30.*
2. *Понкин И. В. Цифровые модели-двойники пациентов: понятие и правовые аспекты // Бизнес, менеджмент и право. 2022. № 2 (54). С. 10–14.*
3. *Иванько А. Ф., Иванько М. А., Гаврилов К. А. IT-технологии обучения и их применение в различных сферах // Молодой ученый. 2019. № 1 (239). С. 5–10.*
4. *Пилявская И. М. Аналитический обзор применения технологий машинного обучения в финансовых ассистентах // Вестник науки и образования. 2022. № 4-2(124). С. 29–34. DOI: 10.24411/2312-8089-2022-10402.*
5. *Сааков Д. В. Применение методов машинного обучения для оптимизации производственных процессов в металлургической промышленности // Инновации и инвестиции. 2023. № 5. С. 308–311.*
6. *Проневич О. Б., Зайцев М. В. Интеллектуальные методы повышения точности прогнозирования редких опасных событий на железнодорожном транспорте // Надежность. 2021. Т. 21. № 3. С. 54–65. DOI: 10.21683/1729-2646-2021-21-3-54-65.*
7. *Энгель Е. А., Энгель Н. Е. Методы машинного обучения для задач прогнозирования и максимизации выработки электроэнергии солнечной электростанции // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. 2023. № 2. С. 146–170. DOI: 10.17308/sait/1995-5499/2023/2/146-170.*
8. *Мочалина М. В., Цапина Т. Н., Чайкина Ж.В . Использование машинного обучения в образовании // Russian Journal of Education and Psychology. 2023. Т. 14. № 1-2. С. 136–140.*
9. *Kotenko I., Saenko I., Lauta O., Kribel K., Vasiliev N. Attacks on artificial intelligence systems: classification, the threat model and the approach to protection // Proceedings of the Sixth International Scientific Conference «Intelligent Information Technologies for Industry» (IITI'22). IITI 2022. Lecture Notes in Networks and Systems, vol 566. Springer, Cham. 2023. Pp. 293–302. DOI: 10.1007/978-3-031-19620-1_28.*
10. *Kotenko I., Saenko I., Lauta O., Vasiliev N., Iatsenko D. Attacks Against Machine Learning Systems: Analysis and GAN-based Approach to Protection // Proceedings of the Seventh International Scientific Conference «Intelligent Information Technologies for Industry» (IITI'23). IITI 2023. Lecture Notes in Networks and Systems, vol 777. Springer, Cham. 2023, pp. 49–59. DOI: 10.1007/978-3-031-43792-2_5.*
11. *Котенко И. В., Саенко И. Б., Лаута О. С., Васильев Н. А., Садовников В.Е. Подход к обнаружению атак на системы машинного обучения с использованием генеративно-состязательной сети // Двадцать первая Национальная конференция по искусственному интеллекту с международным участием, КИИ-2023 (Смоленск, 16-20 октября 2023 г.). Труды конференции. В 2-х томах. Т.1. 2023. С. 366–376.*
12. *Huang X., Kroening D., Ruan W., Sharp J., Sun Y., Thamo E., Wu M., Yi X. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability // Computer Science Review, Volume 37, 2020, 100270. DOI: 10.1016/j.cosrev.2020.100270.*
13. *Martins N., Cruz J. M., Cruz T., Henriques Abreu P., Adversarial Machine Learning Applied to Intrusion and Malware Scenarios: A Systematic Review // IEEE Access, 2020, vol. 8, pp. 35403–35419. DOI: 10.1109/ACCESS.2020.2974752.*
14. *Oseni A., Moustafa N., Janicke H., Liu P., Tari Z., Vasilakos A. Security and Privacy for Artificial Intelligence: Opportunities and Challenges // Journal of ACM, 2020, vol. 37, no. 4, Article 111, 35 pages. DOI: 10.1145/1122445.1122456.*
15. *Xu H., Ma Y., Liu H.C. et al. Adversarial Attacks and Defenses in Images, Graphs and Text: A Review // International Journal of Automation and Computing, 2020, vol. 17, pp. 151–178. DOI: 10.1007/s11633-019-1211-x.*

16. *Ren K., Zheng T., Qin Zh., Liu X. Adversarial Attacks and Defenses in Deep Learning // Engineering, 2020, vol. 6, pp. 346–360. DOI: 10.1016/j.eng.2019.12.012.*
17. *Zhou X., Canady R., Li Y., Koutsoukos X., Gokhale A. Overcoming Stealthy Adversarial Attacks on Power Grid Load Predictions Through Dynamic Data Repair // In: Darema, F., Blasch, E., Ravela, S., Aved, A. (eds) Dynamic Data Driven Applications Systems. DDDAS 2020. Lecture Notes in Computer Science, 2020, vol. 12312, pp 102–109. DOI: 10.1007/978-3-030-61725-7_14.*
18. *Akhtar N., Mian A., Kardan N., Shah M. Advances in Adversarial Attacks and Defenses in Computer Vision: A Survey // IEEE Access, vol. 9, pp. 155161–155196, 2021, DOI: 10.1109/ACCESS.2021.3127960.*
19. *Zhang H., Liu B., H Wu. Smart Grid Cyber-Physical Attack and Defense: A Review // IEEE Access, vol. 9, pp. 29641-29659, 2021, DOI: 10.1109/ACCESS.2021.3058628.*
20. *Chakraborty A., Alam M., Dey V., Chattopadhyay A., Mukhopadhyay D. A survey on adversarial attacks and defences // CAAI Transactions on Intelligence Technology, 2021, vol. 6, pp. 25–45. DOI: org/10.1049/cit2.12028.*
21. *Rosenberg I., Shabtai A., Elovici Y., Rokach L. Adversarial Machine Learning Attacks and Defense Methods in the Cyber Security Domain // ACM Computing Surveys, 2021, vol. 54, no. 5, Article 108, 36 pages. DOI: 10.1145/3453158.*
22. *Tian J., Wang B., Li J., Konstantinou C. Adversarial attack and defense methods for neural network based state estimation in smart grid // IET Renewable Power Generation, 2021, vol. 16, no. 16, pp. 3507-3518. DOI: 10.1049/rpg2.12334.*
23. *Kong Z., Xue J., Wang Y., Huang L., Niu Z., Li F., Meng W. A Survey on Adversarial Attack in the Age of Artificial Intelligence // Wireless Communications and Mobile Computing. 2021. Vol. 2021, Article ID 4907754, 22 pages. DOI: 10.1155/2021/4907754.*
24. *Zhou Sh., Liu Ch., Ye D., Zhu T., Zhou W., Yu Ph.S. Adversarial Attacks and Defenses in Deep Learning: From a Perspective of Cybersecurity // ACM Computing Surveys, 2022, vol. 55, no. 8, Article 163, 39 pages. DOI: 10.1145/3547330.*
25. *Khamaiseh S.Y., Bagagem D., Al-Alaj A., Mancino M., Alomari H.W. Adversarial Deep Learning: A Survey on Adversarial Attacks and Defense Mechanisms on Image Classification // IEEE Access, vol. 10, pp. 102266–102291, 2022, DOI: 10.1109/ACCESS.2022.3208131.*
26. *Liang H., He E., Zhao Y., Jia Z., Li H. Adversarial Attack and Defense: A Survey // Electronics, 2022, vol. 11, 1283. DOI: 10.3390/electronics11081283.*
27. *Tian Q., Zhang S., Mao Sh., Lin Y. Adversarial attacks and defenses for digital communication signals identification // Digital Communications and Networks, 2022, in press. DOI: 10.1016/j.dcan.2022.10.010.*
28. *Anastasiou Th., Karagiorgou S., Petrou P., Papamartzivanos D., Giannetsos Th., Tsirigotaki G., Keizer J. Towards Robustifying Image Classifiers against the Perils of Adversarial Attacks on Artificial Intelligence Systems // Sensors, 2022, vol. 22, 6905. DOI: 10.3390/s22186905.*
29. *Li Y., Cheng M., Hsieh Ch. -J., Lee Th. C. M. A Review of Adversarial Attack and Defense for Classification Methods // The American Statistician, 2022, vol. 76, No. 4, pp. 329–345. DOI: 10.1080/00031305.2021.2006781.*
30. *Tian J., Wang B., Li J., Wang Z. Adversarial Attacks and Defense for CNN Based Power Quality Recognition in Smart Grid // IEEE Transactions on Network Science and Engineering, 2022, vol. 9, no. 2, pp. 807–819. DOI: 10.1109/TNSE.2021.3135565.*
31. *Li H., Namiot D. A Survey of Adversarial Attacks and Defenses for Image Data on Deep Learning // International Journal of Open Information Technologies, 2022, vol. 10, no. 5, pp. 9–16.*
32. *Girdhar M., Hong J., Moore J. Cybersecurity of Autonomous Vehicles: A Systematic Literature Review of Adversarial Attacks and Defense Models // IEEE Open Journal of Vehicular Technology, 2023, vol. 4, pp. 417–437. DOI: 10.1109/OJVT.2023.3265363.*
33. *Goyal Sh., Doddapaneni S., Khapra M. M., Ravindran B. A Survey of Adversarial Defenses and Robustness in NLP // ACM Computing Surveys, 2023, vol. 55, no. 14s, Article 332, 39 pages. DOI: 10.1145/3593042.*
34. *Al-Khassawneh Y. A. A Review of Artificial Intelligence in Security and Privacy: Research Advances, Applications, Opportunities, and Challenges // Indonesian Journal of Science and Technology, 2023, vol. 8, no. 1, pp. 79–96. DOI: 10.17509/IJOST.V8I1.52709.*
35. *Sun L. et al. Adversarial Attack and Defense on Graph Data: A Survey // IEEE Transactions on Knowledge and Data Engineering, 2023, vol. 35, no. 8, pp. 7693–7711. DOI: 10.1109/TKDE.2022.3201243.*