УДК 004.032.26

ОСОБЕННОСТИ ОРГАНИЗАЦИИ АТАК НА НЕЙРОННЫЕ СЕТИ ДЛЯ РАСПОЗНАВАНИЯ ОБРАЗОВ

© 2023 М. А. Лапина[1], Н. В. Ржевская[2], Д. В. Котляров[2], Г. Д. Дюдюн[2]

[1] *доцент кафедры информационной безопасности автоматизированных систем*
e-mail: mlapina@ncfu.ru
[2] *студент*
e-mail: natalia070901@gmail.com

Северо-Кавказский федеральный университет (г. Ставрополь)

В статье рассматривается проблема атак на нейронные сети, используемые для распознавания образов, в контексте информационно-телекоммуникационных систем. Статья подробно описывает состязательные атаки на нейронные сети, которые позволяют обмануть систему путем внесения незаметных искажений или модификаций входных данных. Рассматриваются различные виды атак, включая враждебные атаки, исследуются история исследований и разработки методов атаки и защиты. Приводятся конкретные примеры состязательных атак и обсуждаются их последствия, особенно в областях биометрической аутентификации.

Также статья предлагает различные стратегии защиты от состязательных атак, включая состязательное обучение, маскировку градиента и сжатие функций. Обсуждаются преимущества и ограничения этих стратегий, а также возможность использования генеративных моделей для создания более сложных и реалистичных состязательных примеров. Заключительная часть статьи подчеркивает важность обеспечения безопасности систем машинного обучения и отмечает серьезные последствия атак на нейронные сети, особенно в критических областях, таких как автономные автомобили и медицинские системы. Статья представляет ценную информацию для исследователей и практиков в области информационной безопасности и машинного обучения, помогая понять сложности и вызовы, связанные с организацией атак на нейронные сети, и предлагая стратегии защиты для минимизации рисков и повышения надежности систем распознавания.

Ключевые слова: нейронная сеть, машинное обучение, распознавание образов, искусственный интеллект, алгоритм атак, информационная безопасность, состязательные атаки, вредоносное машинное обучение.

FEATURES OF ORGANIZING ATTACKS ON NEURAL NETWORKS FOR PATTERN RECOGNITION

© 2023 M. A. Lapina[1], N. V. Rzhevskaya[2], D. V. Kotlyarov[2], G. D. Dudyun[2]

[1]*Associate Professor of the Department of Information Security of Automated Systems*
e-mail: mlapina@ncfu.ru
[2]*student of the North Caucasian Federal University*
e-mail: natalia070901@gmail.com

North Caucasian Federal University

This article discusses the issue of attacks on neural networks used for pattern recognition in the context of information and telecommunications systems. The article provides a detailed description of adversarial attacks on neural networks, which allow deceiving the system

by introducing imperceptible distortions or modifications to the input data. Various types of attacks, including adversarial attacks, are examined, and the history of research and development of attack and defense methods is explored. The article presents specific examples of adversarial attacks and discusses their consequences, particularly in the field of biometric authentication. It also proposes various defense strategies against adversarial attacks, including adversarial training, gradient masking, and function compression. The advantages and limitations of these strategies are discussed, as well as the potential use of generative models to create more complex and realistic adversarial examples.

The concluding part of the article emphasizes the importance of ensuring the security of machine learning systems and highlights the serious consequences of attacks on neural networks, especially in critical areas such as autonomous vehicles and medical systems. The article provides valuable information for researchers and practitioners in the fields of information security and machine learning, helping to understand the complexities and challenges associated with organizing attacks on neural networks and offering defense strategies to minimize risks and enhance the reliability of recognition systems.

Keywords: neural network, machine learning, pattern recognition, artificial intelligence, attack algorithm, information security, adversarial attacks, malicious machine learning.

В современное время такие понятия, как «нейронная сеть», «искусственный интеллект» и другие сквозные технологии, используются в разных областях, такие технологии плотно входят в нашу жизнь и уже нередко используются повсеместно. Сейчас никого не удивляет использование поисковых алгоритмов от различных передовых компаний. Даже старшее поколение сейчас часто использует голосовые помощники, не задумываясь, что это еще одно «детище» нейронной сети. В действительности нейросеть сейчас может довольно многое: сгенерировать научные доклады, написать стихи или песню, нарисовать картину, мало отличающуюся от работы настоящего художника, – главное правильно ее обучить.

В настоящий момент сложно однозначно дать определение нейронным сетям. Проанализировав исследование нескольких авторов, можно сказать, что нейронная сеть, или ИНС, – это обучаемая система, которая представляет собой определенную математическую модель, построенную по принципу нейронов человека, а также ее программную реализацию [4; 6]. В.А. Иванюк утверждает, что искусственные нейронные сети могут применяться для создания интеллектуальных систем принятия решений, имитационного моделирования, экспертных систем [4].

К.С. Качагина приводит спектр применения нейронной сети в повседневной жизни. Так, в ее исследовании отмечается, что ИНС уже используется в охранных организациях, правоохранительных органах, на различных заводах и многом другом [6]. Потому очень важно организовать безопасность данных систем, так как дальше область применения нейронных систем будет только расти.

Перечислим основные задачи нейронных сетей:

– классификация, то есть отделение определенного предмета с определенным признаком от других;

– предсказание, эта задача зачастую служит в интересах финансового мира;

– распознавание, помогающее упростить работу, например, правоохранительным органам;

– решение задач без учителя.

В последние годы происходит внедрение нейронных сетей в информационно-телекоммуникационные системы в качестве средства идентификации, а зачастую и аутентификации пользователей [1]. По исследованию экспертов, внедрение подобных технологий часто несет за собой массовые недовольства со стороны. Причиной тому послужило то, что нейросеть не идеальна: в такой системе имеется ряд уязвимостей, который может использоваться для ее выведения из строя [6].

Существует множество атак на нейронные сети, которые препятствуют правильной работе системы. Злоумышленник может реализовывать масштабные атаки, оставаясь незамеченным. Например, в биометрические системы нарушитель может намеренно вносить ошибки в процесс распознавания биометрических данных. Обеспечение безопасности таких систем является важной проблемой.

В статье «Распознавание атаки подделки лица с использованием дискриминационных патчей изображения» (авторы – З. Ахтар, Л. Форести, факультет математики и компьютерных наук, Университет Удине, Италия) описывается, как работают состязательные атаки, используя уязвимости нейронных сетей, которые можно легко обмануть небольшими помехами или модификациями входных данных, незаметными для человека, но приводящими к тому, что сеть неправильно классифицирует входные данные [2].

Рассмотрим некоторые виды атак на биометрические системы, которые нарушают процесс распознавания [1].

1. Fast Gradient Sign Method – атака с наложением шума на изображение с каждой новой итерацией. Эта атака является достаточно эффективной при постоянном анализе изображения. Данный вид атаки практически нереализуем при отсутствии прямого доступа к данным.

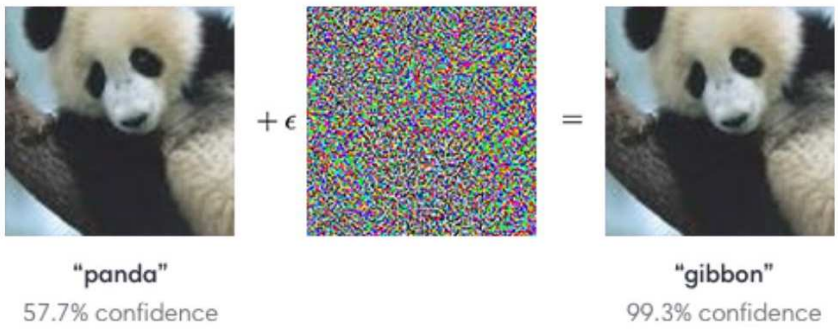

Рис. 1. Пример FGSM

2. Использование инфракрасных светодиодов для изменения черт лица человека.

3. Наложение черных или белых наклеек на изображение для некорректного распознавания.

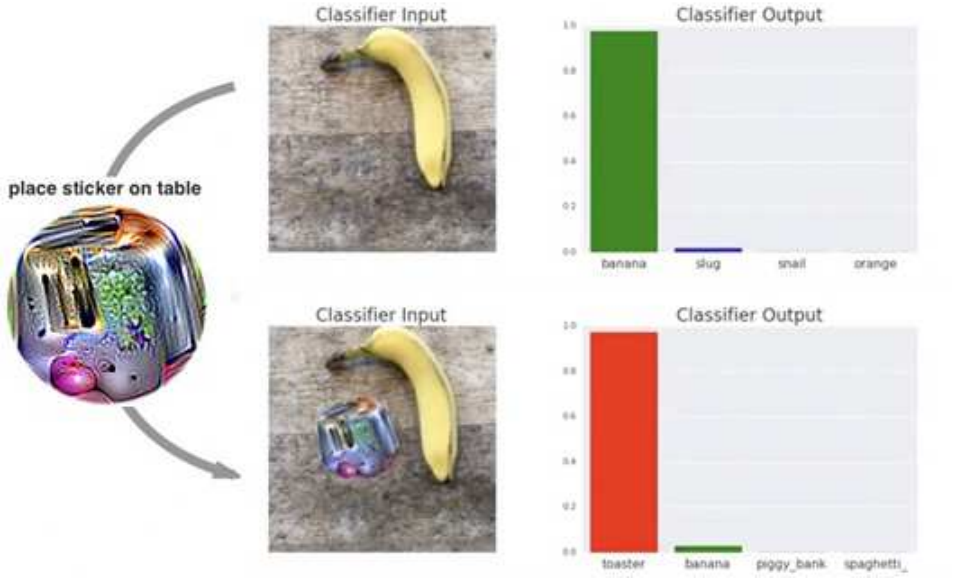

Рис. 2. Применение состязательной наклейки

4. Использование устройств, которые позволяют идентифицировать человека за другого.

Рис. 3. Очки со специально подобранным паттерном

Враждебные атаки вызывают все большую озабоченность в области искусственного интеллекта, поскольку их можно использовать для обмана нейронных сетей и заставлять их неправильно классифицировать входные данные.

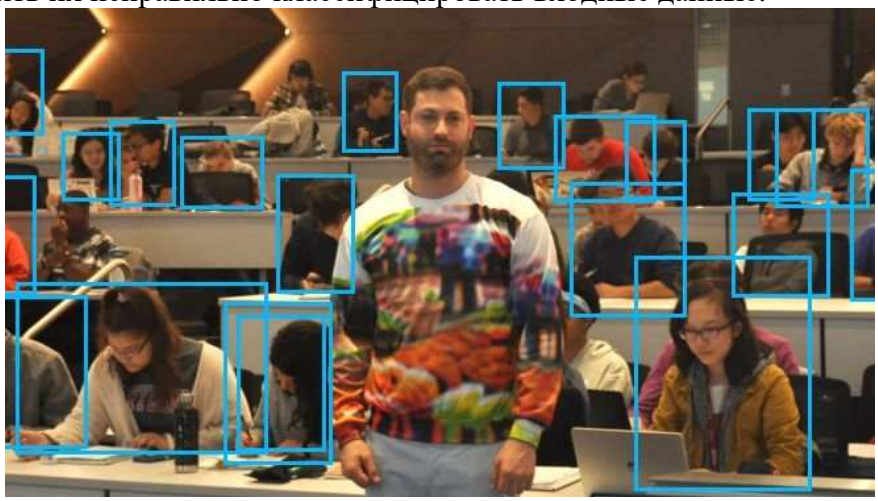

Рис. 4. Узор для одежды, который может обмануть системы распознавания лиц

Одно из первых исследований состязательных атак было проведено Szegedy, оно показало, что нейронные сети можно обмануть небольшими помехами входных данных. С тех пор по этой теме было проведено большое количество исследований, включая разработку новых методов атаки и стратегий защиты.

Одним из наиболее распространенных типов состязательных атак является метод быстрого знака градиента (FGSM), который был представлен Гудфеллоу [3]. Этот метод включает в себя вычисление градиента функции потерь по отношению к входным данным, а затем изменение данных в направлении градиента, чтобы максимизировать потери. Многие последующие исследования основывались на этом методе, в том числе итеративная атака FGSM (IFGSM), представленная Куратин.

В статье «Интерпретация глубоких нейронных сетей…» также объясняются различные типы враждебных атак, в том числе целевые и нецелевые атаки, и приводятся примеры реальных применений враждебных атак, таких как манипулирование системами беспилотных автомобилей [5].

Другие типы атак включают атаку *карты значимости* Jacobian-based, которая использует матрицу Якоби для определения наиболее чувствительных входных признаков, и атаку *deep fool*, которая генерирует небольшие возмущения, минимизирующие расстояние между исходными входными данными и ошибочно классифицированными выходными данными.

Для защиты от этих атак были предложены различные стратегии. Состязательное обучение включает в себя дополнение обучающих данных состязательными примерами, чтобы сделать нейронную сеть более надежной, в то время как защитная дистилляция включает в себя обучение отдельной сети обнаружению состязательных примеров. Другие стратегии защиты включают рандомизацию, преобразование входных данных и градиентную маскировку.

Несмотря на эти стратегии защиты, атаки со стороны противника остаются серьезной угрозой для систем машинного обучения. Как отмечают Ахтар и Намиот, атаки со стороны противника могут иметь серьезные последствия в реальном мире, например, заставить беспилотные автомобили неправильно интерпретировать дорожные знаки или медицинские системы неправильно диагностировать болезни [2; 8].

Кроме того, в статье А Чернобровова обсуждаются некоторые из методов, разработанных для защиты от атак противника, включая обучение противника и защитную дистилляцию [9].

Несмотря на значительный прогресс в разработке состязательных атак и средств защиты, все еще остаются открытые проблемы, требующие дальнейших исследований. Одной из таких задач является разработка действенных и надежных методов защиты. Еще одна проблема заключается в понимании уязвимостей моделей глубокого обучения для атак со стороны злоумышленников и поиске способов их устранения.

В нескольких исследованиях изучалась эффективность состязательных атак на различные типы моделей машинного обучения, включая сверточные нейронные сети (CNN), рекуррентные нейронные сети (RNN) и автоэнкодеры [3; 8; 9].

Для защиты от вражеских атак исследователи предложили различные стратегии защиты. Одной из распространенных стратегий защиты является состязательное обучение, которое включает обучение модели на состязательных примерах в дополнение к обычным обучающим данным [3]. Другие стратегии защиты используют маскировку градиента: сокрытие градиентов от злоумышленника и сжатие функций, в том числе предварительную обработку входных данных для удаления избыточных функций.

В нескольких недавних исследованиях также изучалось использование генеративных моделей, таких как генеративно-состязательные сети (GAN), для создания состязательных примеров. Эти модели можно использовать для создания более реалистичных примеров состязательных действий, которые труднее обнаружить и от которых сложнее защититься.

Системы биометрической аутентификации, которые используют физиологические или поведенческие характеристики для проверки личности людей, становятся все более популярными в последние годы. Однако эти системы не застрахованы от атак [3].

Одним из распространенных типов атак на биометрические данные являются атаки представления, также известные как спуфинговые атаки, когда злоумышленник использует фальшивую или искусственную биометрическую характеристику, чтобы

выдать себя за законного пользователя. Нейронные сети использовались для создания атак с реалистичным представлением различных биометрических модальностей, включая отпечатки пальцев, распознавание лиц и распознавание голоса.

Другие исследования были сосредоточены на использовании нейронных сетей для запуска атак на биометрические данные путем использования уязвимостей в биометрической системе. Исследователи предложили метод создания состязательных примеров для систем распознавания отпечатков пальцев путем искажения входного изображения отпечатка пальца с использованием метода оптимизации на основе градиента. Полученный отпечаток злоумышленника можно использовать, чтобы избежать обнаружения биометрической системой.

Для защиты от атак на биометрические данные исследователи предложили различные стратегии защиты, в том числе использование методов обнаружения живучести для обнаружения атак на презентации и использование глубоких нейронных сетей для повышения устойчивости биометрических систем к атакам. Так, исследователи из Китая предложили метод на основе глубокой нейронной сети для обнаружения презентационных атак в системах распознавания лиц.

Итак, состязательные атаки вызывают особое внимание в области искусственного интеллекта, в этом направлении проведено большое количество исследований. Хотя было предложено немало стратегий, постоянно разрабатываются новые методы нападения. Таким образом, для исследователей важно изучать состязательные атаки и разрабатывать новые стратегии защиты для обеспечения безопасности систем машинного обучения, формирования безопасной модели доверенного искусственного интеллекта.

Библиографический список

1. Атаки на биометрические системы // Безопасность информации. – URL: https://www.itsec.ru/articles/ataka-na-biometricheskie-sistemy (дата обращения: 30.03.2023).

2. *Ахтар, З.* Распознавание атаки подделки лица с использованием дискриминационных патчей изображения / З. Ахтар, Л. Форести; факультет математики и компьютерных наук, Университет Удине, Италия // Journal of Electrical and Computer Engineering. Vol. 2016. – № статьи 4721849. – 14 стр.

3. *Гудфеллоу, И.* Объяснение и использование состязательных примеров / *И. Гудфеллоу, Дж. Шленс, К. Сегеди*; перевод с английского. – URL: https://translated.turbopages.org/proxy_u/en-ru.ru.e1951765-64a3d29c-8c2600eb-74722d776562/https/paperswithcode.com/paper/explaining-and-harnessing-adversarial

4. *Иванюк, В. А.* Нейронные сети и их анализ / В. А. Иванюк // Хроноэкономика. – 2021. – № 4 (32). – URL: https://cyberleninka.ru/article/n/neyronnye-seti-i-ih-analiz (дата обращения: 30.03.2023).

5. Интерпретация глубоких нейронных сетей с помощью SVCCA https://arxiv.org/abs/1706.05806

6. *Качагина, К. С.* NETWORKS NERON – перспективы развития / К. С. Качагина, А. Д. Сафарова // E-Scio. – 2021. – № 2 (53). – URL: https://cyberleninka.ru/article/n/neyronnye-seti-perspektivy-razvitiya (дата обращения: 30.03.2023).

7. *Мамирходжаев, М. Мч.* Возможности нейронных сетей / М. Мч. Мамирходжаев, Д. Т. Умаралиев, М. Б. Сотволдиева, А. Э. Туйчибоев // Талкин ва тадкикотлар илмий-услуги журнали. – 2022. – №6. – URL:

https://cyberleninka.ru/article/n/vozmozhnosti-neyronnyh-setey (дата обращения: 30.03.2023).

8. *Намиот, Д. Э.* Атаки на системы машинного обучения – общие проблемы и методы / Д. Э. Намиот, Э. А. Ильюшин, И. В. Чижов // Международный журнал открытых информационных технологий. – 2022. – №3. – URL: https://cyberleninka.ru/article/n/ataki-na-sistemy-mashinnogo-obucheniya-obschie-problemy-i-methody (дата обращения: 30.03.2023).

9. *Чернобровов, А.* Как обмануть нейросеть, или Что такое состязательная атака / Алексей Чернобровов // АНАЛИТИК. – URL: https://chernobrovov.ru/articles/kak-obmanut-nejroset-ili-chto-takoe-adversarial-attack.html (дата обращения: 02.04.2023).