НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
июль–август 2023 Том 23 № 4 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
July–August 2023 Vol. 23 No 4 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

doi: 10.17586/2226-1494-2023-23-4-720-733
УДК 004.056

Атаки на основе вредоносных возмущений на системы обработки изображений и методы защиты от них

Дмитрий Андреевич Есипов[1]✉, Абдулхамид Яхьяевич Бучаев[2], Акылжан Керимбай[3], Яна Владиславовна Пузикова[4], Семен Кириллович Сайдумаров[5], Никита Сергеевич Сулименко[6], Илья Юрьевич Попов[7], Николай Сергеевич Кармановский[8]

[1,2,3,4,5,6,7,8] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

[1] some1else.d.ma@gmail.com✉, https://orcid.org/0000-0003-4467-5117
[2] abdulhamid0055@yandex.ru, https://orcid.org/0009-0001-1058-9125
[3] akerimbai@itmo.ru, https://orcid.org/0009-0009-9945-9906
[4] Yanapuzikova19@ya.ru, https://orcid.org/0009-0007-7604-3022
[5] semen.say@ya.ru, https://orcid.org/0009-0008-0774-9803
[6] n.s.sulimenko@mail.ru, https://orcid.org/0009-0007-3218-9249
[7] ilyapopov27@gmail.com, https://orcid.org/0000-0002-6407-7934
[8] karmanov50@mail.ru, https://orcid.org/0000-0002-0533-9893

Аннотация

Системы, реализующие технологии искусственного интеллекта, получили широкое распространение благодаря их эффективности в решении прикладных задач, включая компьютерное зрение. Обработка изображений посредством нейронных сетей применяется в критически важных для безопасности системах. В то же время использование искусственного интеллекта сопряжено с характерными угрозами, к которым относится и нарушение работы моделей машинного обучения. Феномен провокации некорректного отклика нейронной сети посредством внесения визуально незаметных человеку искажений впервые описан и привлек внимание исследователей в 2013 году. Методы атак на нейронные сети на основе вредоносных возмущений непрерывно совершенствовались, были предложены способы нарушения работы нейронных сетей при обработке различных типов данных и задач целевой модели. Угрозы нарушения функционирования нейронных сетей посредством указанных атак стала значимой проблемой для систем, реализующих технологии искусственного интеллекта. Таким образом, исследования в области противодействия атакам на основе вредоносных возмущений являются весьма актуальными. В данной статье представлено описание актуальных атак, приведен обзор и сравнительный анализ таких атак на системы обработки изображений с использованием искусственного интеллекта. Сформулированы подходы к классификации атак на основе вредоносных возмущений. Рассмотрены методы защиты от подобных атак, выявлены их недостатки. Показаны ограничения применяемых методов защиты, снижающие эффективность противодействия атакам. Предложены подходы по обнаружению и устранению вредоносных возмущений.

Ключевые слова

искусственный интеллект, искусственная нейронная сеть, обработка изображений, состязательная атака, встраивание бэкдора, вредоносное возмущение, состязательное обучение, защитная дистилляция, сжатие параметров, сертификационная защита, предобработка данных

Ссылка для цитирования: Есипов Д.А., Бучаев А.Я., Керимбай А., Пузикова Я.В., Сайдумаров С.К., Сулименко Н.С., Попов И.Ю., Кармановский Н.С. Атаки на основе вредоносных возмущений на системы обработки изображений и методы защиты от них // Научно-технический вестник информационных технологий, механики и оптики. 2023. Т. 23, № 4. С. 720–733. doi: 10.17586/2226-1494-2023-23-4-720-733

Attacks based on malicious perturbations on image processing systems and defense methods against them

Dmitry A. Esipov[1]✉, Abdulhamid Y. Buchaev[2], Akylzhan Kerimbay[3], Yana V. Puzikova[4], Semen K. Saidumarov[5], Nikita S. Sulimenko[6], Ilya Yu. Popov[7], Nikolay S. Karmanovskiy[8]

[1,2,3,4,5,6,7,8] ITMO University, Saint Petersburg, 197101, Russian Federation

[1] some1else.d.ma@gmail.com✉, https://orcid.org/0000-0003-4467-5117
[2] abdulhamid0055@yandex.ru, https://orcid.org/0009-0001-1058-9125
[3] akerimbai@itmo.ru, https://orcid.org/0009-0009-9945-9906
[4] Yanapuzikova19@ya.ru, https://orcid.org/0009-0007-7604-3022
[5] semen.say@ya.ru, https://orcid.org/0009-0008-0774-9803
[6] n.s.sulimenko@mail.ru, https://orcid.org/0009-0007-3218-9249
[7] ilyapopov27@gmail.com, https://orcid.org/0000-0002-6407-7934
[8] karmanov50@mail.ru, https://orcid.org/0000-0002-0533-9893

Abstract

Systems implementing artificial intelligence technologies have become widespread due to their effectiveness in solving various applied tasks including computer vision. Image processing through neural networks is also used in security-critical systems. At the same time, the use of artificial intelligence is associated with characteristic threats including disruption of machine learning models. The phenomenon of triggering an incorrect neural network response by introducing perturbations that are visually imperceptible to a person was first described and attracted the attention of researchers in 2013. Methods of attacks on neural networks based on malicious perturbations have been continuously improved, ways of disrupting the operation of neural networks in processing various types of data and tasks of the target model have been proposed. The threat of disrupting the functioning of neural networks through these attacks has become a significant problem for systems implementing artificial intelligence technologies. Thus, research in the field of countering attacks based on malicious perturbations is very relevant. This article describes current attacks, provides an overview and comparative analysis of such attacks on image processing systems based on artificial intelligence. Approaches to the classification of attacks based on malicious perturbations are formulated. Defense methods against such attacks are considered, their shortcomings are revealed. The limitations of the applied defense methods that reduce the effectiveness of counteraction to attacks are shown. Approaches and practical measures to detect and eliminate harmful disturbances are proposed.

Keywords

artificial intelligence, artificial neural network, image processing, adversarial attack, backdoor embedding, adversarial perturbation, adversarial learning, defense distillation, feature squeezing, certified defense, data preprocessing

For citation: Esipov D.A., Buchaev A.Y., Kerimbay A., Puzikova Y.V., Saidumarov S.K., Sulimenko N.S., Popov I.Yu., Karmanovskiy N.S. Attacks based on malicious perturbations on image processing systems and defense methods against them. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2023, vol. 23, no. 4, pp. 720–733 (in Russian). doi: 10.17586/2226-1494-2023-23-4-720-733

Введение

Системы, реализующие технологии искусственного интеллекта, получили широкое распространение за счет значительной эффективности по сравнению с другими методами при решении множества прикладных задач, в том числе в обработке естественного языка [1], распознавании речи [2], биометрической аутентификации [3–5], медицинской диагностике [6, 7], видеонаблюдении [8], беспилотном транспорте [9–11] и др. Указанные системы нередко используются для обработки изображений в критической информационной инфраструктуре. По прогнозам компании Gartner, количество беспилотных транспортных средств увеличится на 740 тысяч единиц по всему миру[1].

В то же время использование систем, реализующих технологии искусственного интеллекта, и нейронных сетей сопряжено с характерными угрозами безопасности[2,3]. Значительную опасность представляют такие угрозы, как нарушение функционирования или обход средств, реализующих технологии искусственного интеллекта, и модификации модели машинного обучения при помощи искажения данных. За 2022 год 30 % кибератак были направлены на обход или кражу моделей нейронных сетей [12]. Такие угрозы могут быть реализованы в виде атак посредством внесение вредоносных возмущений [13–69]. Отметим, что атаки могут быть реализованы вне зависимости от типа обрабатываемых данных [13]. Наибольшее распространение указанные атаки получили для систем обработки изображений и видеопотоков.

Состязательные атаки приводят к некорректному отклику нейронных сетей и, как следствие, нарушению работы системы, реализующей технологии искус-

[1] Rimol M. Gartner Forecasts More Than 740,000 Autonomous-Ready Vehicles to Be Added to Global Market in 2023 [Электронный ресурс]. 2019. Режим доступа: https://www.gartner.com/en/newsroom/press-releases/2019-11-14-gartner-forecasts-more-than-740000-autonomous-ready-vehicles-to-be-added-to-global-market-in-2023, свободный (дата обращения: 21.01.2023).

[2] ФСТЭК. Банк данных угроз безопасности информации [Электронный ресурс]. Режим доступа: https://bdu.fstec.ru/threat, свободный (дата обращения: 03.03.2023).

[3] MITRE. Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS) [Электронный ресурс]. Режим доступа: https://atlas.mitre.org/, свободный (дата обращения: 03.03.2023).

ственного интеллекта, что может повлечь критические последствия в зависимости от области применения: некорректная постановка диагноза, игнорирование дорожных знаков беспилотным автомобилем и т. д. Кроме того, модификация модели машинного обучения может приводить к манипулированию ее откликами, провоцируя желаемое для злоумышленника поведение [14].

В настоящей работе рассмотрены атаки на основе вредоносных возмущений, их формальное описание, подходы к классификации и конкретные методы атаки, а также методы защиты от указанных атак, их достоинства и недостатки. Кроме того, предложены подходы к обнаружению и устранению вредоносных искажений, встраиваемых в изображения при атаке.

Атаки на основе вредоносных возмущений

Атаки на основе вредоносных возмущений предполагают внесение искажений во входные данные, приводящих к нарушению функционирования целевой системы. Нормой таких возмущений являются метрики расстояния [13], характеризующие вредоносное возмущение как разницу между оригинальным и искаженным изображениями. Для них характерны следующие аксиомы:

— неотрицательность: $||x|| \geq 0$;
— тождество неразличимых: $||x - y|| = 0 \leftrightarrow x = y$;
— симметричность: $||x - y|| = ||y - x||$;
— неравенство треугольника: $||x - y|| + ||y - z|| > ||x - z||$.

На практике активно используются следующие метрики (рис. 1):

— L_1 (манхэттенское расстояние):

$$||x||_1 = \sum_{i=1}^{n} |x_i|,$$

где x — вредоносное возмущение; n — число пикселов; x_i — i-й пиксел;

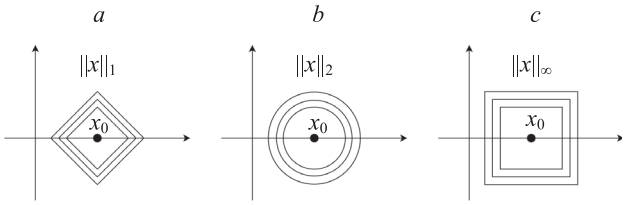

Рис. 1. Метрики расстояния: манхэттенское расстояние L_1 (*a*); евклидова метрика L_2 (*b*); расстояние Чебышева L_∞ (*c*)

Fig. 1. Distance metrics: Manhattan distance L_1 (*a*); Euclid distance L_2 (*b*); Chebyshev distance L_∞ (*c*)

— L_2 (евклидова метрика):

$$||x||_2 = \sqrt{\sum_{i=1}^{n} x_i^2};$$

— L_∞ (расстояние Чебышева):

$$||x||_\infty = \max_i |x_i|.$$

В дополнение к перечисленным метрикам также используется метрика L_0: $||x||_0 = |\{x_i | x_i \neq 0, i = 1, \ldots, n\}|$, показывающая число ненулевых элементов (в случае изображений — модифицированных пикселов). Указанная метрика не соответствует аксиомам, потому формально не является метрикой расстояния и нормой возмущения. L_0 также называют псевдонормой.

Классификации атак

Атаки на основе внесения вредоносных возмущений на модели машинного обучения могут быть классифицированы по следующим признакам: преследуемая цель; тип обрабатываемых данных; задача целевой модели; тип искажения; знание атакующего о целевой системе; направленность; число преобразований (рис. 2).

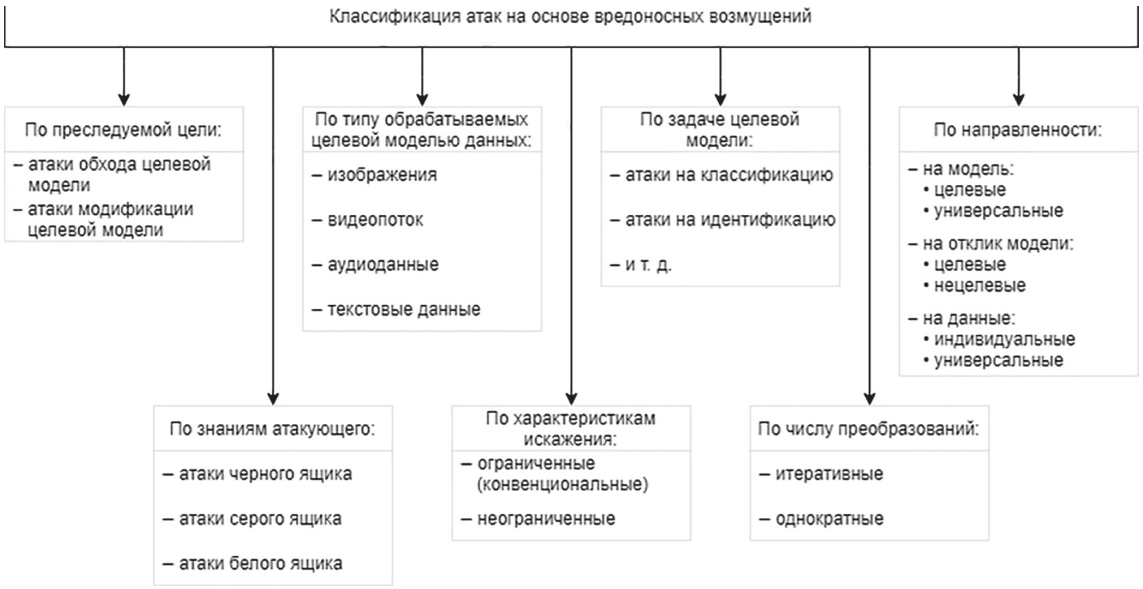

Рис. 2. Классификация атак на основе вредоносных возмущений

Fig. 2. Classification of attacks based on malicious perturbation

Важно отметить, что атаки на основе вредоносных возмущений характерны в первую очередь для искусственных глубоких нейронных сетей (Deep Neural Network, DNN), в то же время, они также могут быть направлены и на другие технологии машинного обучения. В текущей работе будут рассмотрены атаки на нейронные сети.

Классификация по преследуемой цели. По преследуемой цели можно выделить следующие типы атак:
— атака обхода целевой модели;
— атака модификации целевой модели.

Атака обхода (уклонения, evasion) соответствует реализации угрозы УБИ.220 «Угроза нарушения функционирования («обхода») средств, реализующих технологии искусственного интеллекта» согласно банку данных угроз ФСТЭК, тогда как атака модификации целевой модели — УБИ.221 «Угроза модификации модели машинного обучения путем искажения («отравления») обучающих данных»[1]. Аналогичные техники обозначены также и в базе знаний MITRE ATLAS: Evade ML Model и Backdoor ML Model[2] соответственно.

$$N(x) = y;\ N(x + p) = y';\ y \neq y', \tag{1}$$

где N — нейронная сеть; x — входные данные; y — легитимный отклик модели; p — вредоносное искажение, y' — нелегитимный отклик модели.

Атаки обхода целевой модели, также известные как состязательные атаки (adversarial attack) [15–65], предполагают модификацию входных данных для изменения отклика модели и поведения системы, реализующей технологии искусственного интеллекта (1). Такие атаки не оказывают влияния на целостность целевой модели.

$$S'\colon \forall x' \in S',\ x' = x + p_t;\ L(N_0,\ S + S') = N', \tag{2}$$

где S — обучающая выборка без искаженных данных; S' — набор искаженных данных; x' — искаженный экземпляр; p_t — встраиваемый триггер; $L(\cdot)$ — процедура обучения модели; N_0 — подготовленная к обучению модель; N' — модифицированная нейронная сеть.

$$N'(x) = y;\ N'(x + p_t) = y_t';\ y \neq y_t', \tag{3}$$

где y_t' — некорректный отклик модели, провоцируемый триггером p_t.

Атаки модификации целевой модели [66–69], основанные на вредоносных возмущениях, предполагают в первую очередь модификацию обучающей выборки целевой нейронной сети посредством модификации или добавления искаженных элементов и обучения модели на этих данных (2). В случае встраивания бэкдора, вредоносное возмущение представляет собой триггер, провоцирующий конкретный отклик модели.

Дальнейшее использование бэкдора предполагает наложение возмущения, идентичного триггеру, на входные данные (3).

В текущей работе преимущественно будут рассмотрены состязательные атаки.

Классификация по типу обрабатываемых данных. Атаки также могут быть классифицированы в зависимости от типа обрабатываемых целевой моделью данных:
— изображения [15–59, 66–70];
— видеопоток [60, 61];
— аудиоданные [62, 63];
— текстовые данные [64, 65].

Следует отметить, что к текстовым данным также относится и сетевой трафик [64]. В текущей работе будут рассмотрены атаки на системы обработки изображений.

Классификация по задаче целевой модели. По задаче целевой модели можно выделить следующие типы атак:
— атаки на системы классификации данных [15–56, 60–69];
— сегментация изображения [57, 58];
— повышение разрешающей способности [59];
— и др.

Потенциально, к атакам на основе вредоносных возмущений могут быть уязвимы системы, реализующие технологии искусственного интеллекта, вне зависимости от решаемых задач.

Классификация по характеристикам вносимого возмущения. По характеристикам искажения можно выделить следующие типы атак [16]:
— атаки с ограниченным возмущением (restricted perturbation);
— атаки с неограниченным возмущением (unrestricted perturbation).

$$N(x) = y;\ N(x + p) = y';\ y \neq y';\ \|p\|_p \leq \eta, \tag{4}$$

где $\|p\|_p$ — норма вредоносного искажения; η — некоторая константа.

Атаки с ограниченным возмущением предполагают ограничение возмущения некоторым значением метрики расстояния . Такие атаки обхода целевой модели принято называть конвенциональными состязательными атаками (conventional adversarial attack) [16] (4). Ограничение на вносимые возмущения позволяет им оставаться невидимыми или слабо заметными для человека, однако оказывать значительное влияние на отклик нейронной сети [15].

Следует отметить, что вредоносное возмущение может быть ограничено как в атаках обхода [15–50], так и в атаках модификации [66] целевой модели.

Атаки с неограниченным возмущением предполагают произвольную модификацию входных данных вплоть до их подмены [51–56, 67–69]. К указанным искажениям относятся управление цветом [51], поворот изображения [52], управление атрибутами [53], наложение состязательных патчей [52] и не только.

[1] ФСТЭК. Банк данных угроз безопасности информации [Электронный ресурс]. Режим доступа: https://bdu.fstec.ru/threat, свободный (дата обращения: 03.03.2023).

[2] MITRE. Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS) [Электронный ресурс]. Режим доступа: https://atlas.mitre.org/, свободный (дата обращения: 03.03.2023).

К таким атакам в том числе можно отнести и создание дипфейков (DeepFake) [55, 56].

Классификация по знаниям атакующего о целевой модели. В зависимости от известной атакующему информации о целевой системе, атаки на основе внесения вредоносных возмущений можно выделить следующие типы атак:

— атаки по модели белого ящика (white-box attack);
— атаки по модели черного ящика (black-box attack).

В случае атак по модели белого ящика [15, 17–32, 34] атакующему доступна вся информация о целевой системе, в том числе архитектура и параметры нейронной сети. Например, атаки, использующие знание гессиана и градиента функции потерь.

Атаки по модели черного ящика [33, 35–50] являются более приближенными к реальным ситуациям, так как атакующий не владеет данными об атакуемой системе. Такие атаки могут быть основаны: на отклике целевой модели [35–44] и на переносе вредоносного возмущения [45–50]. Оба типа атак для генерации искажения используют отклик целевой модели, однако атаки, основанные на переносе, в качестве базиса для генерации используют заранее подготовленное возмущение, в том числе может быть использовано универсальное возмущение [34].

В отдельных случаях выделяют атаки по модели серого ящика (gray-box attack), при которых атакующему известна лишь некоторая часть информации о целевой системе. Такие атаки нередко приравнивают к другим типам по рассматриваемой классификации, чаще — к атакам по модели черного ящика.

Классификация по направленности. По направленности можно выделить атаки: целевые (targeted) [15–19, 24, 32, 33, 35–50] и нецелевые [29–31, 34] (non-targeted) по отклику. Целевые по отклику предполагают внесение искажения, провоцирующего конкретный отклик целевой модели, а нецелевые по отклику — провокацию произвольного некорректного отклика.

Также по направленности может быть предложена другая классификация атак: целевые по модели и универсальные (universal) [34]. В этом случае целевые атаки предполагают внесение искажений, провоцирующих некорректный отклик конкретной целевой модели машинного обучения. Универсальные атаки предполагают создание таких искажений, что атака посредством их встраивания позволит вызвать некорректный отклик произвольной модели соответствующего функционала.

Аналогичная классификация может быть предложена и по влиянию вносимого возмущения на входные данные модели: индивидуальные (целевые) [70] по входным данным и универсальные атаки. Первые предполагают нарушение работы целевой системы на конкретном экземпляре входных данных, вторые сохраняют эффект для некоторого множества входных данных.

Классификация по числу преобразований. По числу преобразований можно выделить методы генерации искажения за один шаг (one-step method) [18] и итеративные методы (iterative method) [15–17, 21, 23, 29–31, 35–50]. Следует отметить, что создание вредоносного возмущения за одно преобразование более характерно для атак по модели белого ящика.

Существующие методы атаки на основе вредоносных возмущений на нейронные сети обработки изображений

Исторически первой состязательной атакой на нейронные сети обработки изображения стала L-BFGS (Limited Memory Broyden–Fletcher–Goldfarb–Shanno) [15, 17], предполагавшая итеративную генерацию вредоносного возмущения посредством анализа гессиана функции потерь целевой модели и обратного движения в сторону роста значения функции потерь.

Другим методом атаки на основе вредоносного возмущения, оказавшим значительное влияние на дальнейшее развитие состязательных атак, стал FGSM (Fast Gradient Sign Method) [18, 19]. В основе указанного метода лежит использование знака градиента функции потерь целевой модели и генерация искажения в результате одноэтапного движения по градиенту к метке определенного класса. Рассматриваемый метод имеет меньшее время поиска вредоносного возмущения, однако генерируемое возмущение обладает высоким значением нормы. При этом FGSM характеризуется сравнительно низкой вероятностью успеха.

Многие последующие алгоритмы атак использовали FGSM в качестве основы и фактически являлись модификацией рассмотренного метода, в том числе FGVM (Fast Gradient Value Method) [20] и I-FGSM (Iterative Fast Gradient Sign Method) [21]. Атака по методу FGVM, что следует из названия, предполагает использование значения градиента вместо его знака. Метода атаки I-FGSM в свою очередь предусматривает генерацию искажения итеративно. В работах [22] и [23] предлагается ограничение генерируемого возмущения по метрикам L_2 и L_∞ соответственно. Важно отметить, что атака по методу BIM (Basic Iterative Method) [23] была успешно воспроизведена и на физических объектах. Метод атаки PGD (Projected Gradient Descent) [24] также представляет собой вариацию итеративного FGSM, где для решения задачи оптимизации и нахождения оптимального искажения используется проецирование градиентного спуска. Существуют также и другие методы атаки, основанные на FGSM, в том числе его модификации [25–28].

Метод атаки DeepFool [29] предполагает итеративную генерацию атакующего изображения при помощи выпрямления границ принятия решения целевой модели и аппроксимации по ряду Тейлора для нахождения ближайшего класса к данному до тех пор, пока не будет сгенерировано наиболее подходящее минимальное вредоносное возмущение. Рассматриваемый метод характеризуется высокой вероятностью обхода целевой модели при использовании возмущения малой нормы, однако большое количество итераций приводит к большим вычислительным затратам и росту времени генерации.

Метод C&W (Carlini&Wagner) [30] представляет собой модификацию атаки L-BFGS с заменой нелинейных граничных условий для упрощения решения

задачи оптимизации. Преимуществами рассматриваемой атаки являются скорость генерации и минимизация вредоносного возмущения. Для указанного метода характерно малое смещение вероятностей принадлежности к определенному классу, что позволяет противостоять методам защиты, например методу дистилляции (Defense Distillation) [71].

Метод атаки TR (Trust Region based) [31] предполагает использование доверительных областей [72] для минимизации и ограничения вносимого искажения. Генерация вредоносного возмущения по указанному методу заключается в итеративном вычислении доверительного радиуса для максимизации вероятности некорректного класса внутри доверительной области. Рассмотренный метод достигает сравнимых с C&W показателей по метрикам расстояния для генерируемого возмущения, вследствие чего также способен противостоять методу дистилляции.

В отличие от рассмотренных конвенциональных атак, JSMA (Jacobian-based Saliency Map Attack) [32] и однопиксельная атака (One-pixel attack, Singlepixel attack) [33] вместо ограничения возмущения по метрике расстояния используют ограничение по размеру искажения.

Так, JSMA вычисляет Якобиан (Jacobian) [73] функции целевой модели и по полученным значениям формирует карту значимости (salience map) [74] пикселов изображения. В соответствии с построенной картой значимости определяются элементы, оказывающие наибольшее влияние на отклик модели. Особенностью рассмотренного метода является получение показателей влияния на отклик модели посредством перебора большого числа входных параметров, что приводит к росту времени генерации вредоносного возмущения.

Однопиксельная атака [33] предполагает формирование искажения, не превышающего один пиксел. Для определения положения и значений RGB-компонент пиксела применяется алгоритм дифференциальной эволюции (Differential Evolution Algorithm, DEA) [75], что позволяет организовать атаку по модели черного ящика. Рассмотренный метод атаки более эффективен для небольших изображений [33].

Рассмотренные выше атаки относятся к атакам по модели белого ящика, их сравнение приведено в табл. 1. Несмотря на то, что однопиксельная атака соответствует модели черного ящика, для оценки и последующего сравнения указанного метода более актуальны критерии атак по модели белого ящика, потому он также включен в табл. 1. В качестве критериев сравнения для таких атак выбраны размер и норма искажения, а также доля успешных обходов целевой модели. Размер искажения определен относительно исходного изображения размерами $w \times h$, где w и h — ширина и высота изображения соответственно. В качестве нормы искажения рассмотрены метрики, которые были учтены при генерации вредоносного возмущения при адресации атаки.

Рассмотренные методы генерации атакующих изображений имеют как преимущества, так и недостатки. Генерация вредоносного возмущения минимальной нормы требует больших затрат, однако такие искажения позволяют оставаться незаметными в том числе и для некоторых механизмов защиты [71]. Наиболее эффективным методом генерации вредоносного возмущения, согласно выбранным критериям, является метод TR [31], так как в его основе лежит оптимизация вычисления возмущения наименьшей нормы.

Рассмотрим атаки по модели черного ящика [33, 35–50].

Метод атаки Boundary attack [35] основан на отклике целевой модели. В качестве начального искажения для алгоритма генерации вредоносного возмущения выступает изображение целевого класса. С каждой итерацией алгоритм оптимизирует и уменьшает возмущение до тех пор, пока исходное изображение не будет корректно классифицировано. Тогда используется искажение, полученное на предыдущей итерации алгоритма.

Метод атаки HopSkipJump [36] является развитием Boundary attack и начальное поведение алгоритма генерации возмущения совпадает с предыдущим методом. Однако вблизи границы принятия решения HopSkipJump модифицирует искажение в зависимости от направления градиента. Длина шага вдоль направления градиента уменьшается в геометрической прогрессии для обеспечения минимального значения метрики расстояния возмущения.

Метод qFool [37] минимизирует вносимое искажение посредством нахождения направления градиен-

Таблица 1. Сравнение методов атаки по модели белого ящика

Table 1. Comparison of white-box attacks

Метод	Тип атаки	Размер искажения	Ограничение нормы искажения	Доля успешных атак, %
LBFGS [17]	Конвенциональная итеративная целевая по модели атака обхода	$w \times h$	—	87,65
FGSM [19]		$w \times h$	$\|\|x\|\|_2 \le 6{,}25$	54,33
PGD [24]		$w \times h$	$\|\|x\|\|_\infty \le 0{,}20$	91,85
DeepFool [29]	Конвенциональная итеративная целевая по модели и нецелевая по отклику атака обхода	$\le w \times h$	$\|\|x\|\|_\infty \le 0{,}20$	92,60
C&W [30]		$w \times h$	$\|\|x\|\|_\infty \le 0{,}20$	94,80
TR [31]		$w \times h$	$\|\|x\|\|_\infty \le 0{,}10$	94,77
JSMA [32]	Конвенциональная итеративная целевая по модели атака обхода	$< w \times h$	—	97,05
One-pixel [33]		1×1	$\|\|x\|\|_0 = 1$	72,85

Таблица 2. Сравнение методов атаки по модели черного ящика [38]
Table 2. Comparison of attack methods of black-box attacks

Метод	Тип атаки	L_2	L_∞	Число запросов
Boundary attack	Конвенциональная итеративная целевая по модели и целевая по классу атака обхода	5,13	0,052	800
HopSkipJump		1,85	0,012	42
qFool		1,12	—	3
GeoDA		1,01	0,003	14

та, имеющего кратчайший путь до границы принятия решения. Однако указанная граница нередко имеет значительную степень кривизны, тогда оптимизация процедуры вычисления минимального вредоносного возмущения может быть выполнена с помощью определения оптимальной нормали, что реализовано в методе GeoDA [38].

В работах [39–44] рассмотрены другие методы атаки по модели черного ящика, предполагающие использование отклика целевой модели.

Поскольку методы атаки по модели черного ящика выполняют минимизацию вносимого искажения при сохранении некорректного отклика модели, в качестве критериев сравнения выбраны метрики расстояния и количество итераций преобразования. Сравнение атак приведено в табл. 2.

В результате сравнения видно, что методы qFool [37] и GeoDA [38] достигают наименьших по норме возмущений за меньшее число итераций относительно существующих аналогов. Наиболее эффективным методом атаки по модели черного ящика является GeoDA [38], так как указанный метод решает проблему кривизны геометрической границы принятия решений в общем случае.

Рассмотренные атаки по модели черного ящика использовали только отклик целевой модели для генерации искажения. В то же время для уменьшения числа итераций в качестве основы для вредоносного возмущения могут быть использованы заранее подготовленные искажения. Прежде чем рассматривать такие атаки, необходимо упомянуть универсальные искажения.

В работе [34] были представлены вредоносные возмущения, не зависящие от задачи, решаемой целевой системой, и позволяющие осуществить обход таких систем. Отметим, что такие искажения могут быть перенесены как между различными входными данными, так и нейронными сетями.

Атака Customized Adversarial Boundary (CAB) [45] является развитием Boundary attack. Уменьшение числа запросов для генерации искажения достигается за счет инициализации возмущения, характерного для атак переноса вредоносного искажения, а также вычисления статистического распределения шума в предыдущих запросах.

Метод TRansferable EMbedding based Black-box Attack (TREMBA) [46] предполагает использование модели архитектуры Encoder-Decoder [76] для ограничения пространства поиска вредоносного возмущения, что позволяет значительно ускорять процесс генерации искажения.

В работе [47] предложено использование дополнительной замещающей модели для генерации начального возмущения. Идея метода заключается в применении генератора для синтеза изображений, впоследствии обрабатываемых целевой моделью, и обучении модели на синтезированных данных и полученных откликах. Такой подход позволяет воспроизвести границы принятия решения целевой модели и использовать синтезированные изображения в качестве исходных данных для создания вредоносного возмущения.

Многие исследования развивают идею атак на основе переноса вредоносного возмущения [48–50].

Кроме рассмотренных конвенциональных состязательных атак, также существуют методы, предполагающие неограниченное возмущение [51–56]. Для нарушения работы целевой системы при таких атаках могут быть использованы такие искажения, как поворот изображения [52], значительное изменение цветовых компонент [51], добавление атрибутов [53] и др. К указанным атакам относятся также дипфейки [55, 56], позволяющие подменять человека на изображении или в видеопотоке на другое лицо, формируя подделку.

Другой вид атак на основе вредоносных возмущений — модификация модели посредством встраивания бэкдора [66–69]. Внедрение бэкдора в целевую систему возможно путем добавления триггеров в виде специфичных паттернов возмущений на элементы обучающей выборки. При этом могут быть использованы как ограниченные [66], так и неограниченные возмущения, в том числе в виде состязательных патчей [67].

Метод модификации целевой системы DPatch [67] предполагает нанесение на изображения небольшого видимого искажения, представляющего собой квадратную область размером от 20 × 20 пикселов. Указанный метод ориентирован на системы обнаружения объектов на изображении. Основным недостатком состязательных патчей является их заметность для человека, что значительно упрощает их обнаружение. Для придания триггерам более естественного вида предложены методы, адаптирующие патчи под естественные отражения или тени [68] и использующие плавную незначительную деформацию изображения [69]. В работе [66] разработан метод, предполагающий встраивание бэкдора посредством внесения вредоносного возмущения, сравнимого с конвенциональными состязательными атаками.

Рассмотренные методы атаки в основном направлены на некорректный отклик модели при классификации, однако также существуют методы, направленные на нарушение работы нейронных сетей при решении других задач обработки изображений [57–59].

Методы противодействия атакам на основе вредоносных возмущений

Вследствие актуальности и значимости рассмотренных атак для многих систем, были разработаны методы противодействия им. Существующие методы защиты включают:

— методы модификации целевой модели [17, 77, 78];
— методы модификации целевой системы, реализующей технологии искусственного интеллекта [71, 79, 80];
— предобработку или изменение входных данных [79–83].

Кроме того, некоторые методы предполагают как обнаружение, так и устранение искажения или его влияния на нейронную сеть [17, 77, 78, 80–83], тогда как другие — только обнаружение [71, 79].

Отметим, что методы защиты должны удовлетворять следующим критериям [71]:

— минимальное воздействие на архитектуру;
— минимальное влияние на показатели качества;
— минимальное влияние на быстродействие.

Одним из методов, предполагающих модификацию модели, является состязательное обучение (Adversarial Learning) [17, 77, 79]. Сущность метода заключается в обучении целевой модели в том числе на экземплярах, содержащих вредоносное искажение, однако сопоставляемых с корректным откликом. Согласно исследованиям, при состязательном обучении доля некорректных откликов при адресации атаки, основанной на FGSM, снизилась с 89,4 % до 17,9 % [78]. Однако применение указанного метода защиты влияет и на обработку данных, не содержащих вредоносного возмущения — точность классификации снижается. Сложностью при состязательном обучении является обеспечение репрезентативности обучающей выборки и контроль ее статистического распределения. Кроме того, указанный метод предполагает переобучение целевой нейронной сети, что не всегда возможно.

Такой метод, как защитная дистилляция (Defense Distillation) [71], предполагает обучение дополнительной нейронной сети идентичной архитектуры, однако использующей при обучении в качестве предикторов отклики защищаемой модели, представляющие собой вероятности отнесения входного изображения к тому или иному классу. При генерации и наложении вредоносного искажения вероятность отнесения входных данных к некорректному классу нередко достигает значений, не характерных для «чистых» изображений. Такая уверенность в классификации представляет собой аномалию, которая может быть обнаружена защищающей нейронной сетью. Однако защитная дистилляция не позволяет обнаруживать некоторые виды атак, например, C&W [30].

Другим вариантом противодействия рассмотренным атакам является сжатие параметров входных данных (Feature Squeezing) [79], что может быть достигнуто в том числе при понижении размерности метаданных. Метод предполагает обработку как оригинального изображения моделью машинного обучения защищаемой системы, реализующей технологии искусственного интеллекта, так и модифицированного изображения дополнительной нейронной сетью. Модификация входных данных нарушает целостность вредоносного возмущения, вносимого атакой, в результате чего нейронные сети вернут различные отклики, что будет свидетельствовать о факте атаки. Сжатие параметров и защитная дистилляция предполагают использование дополнительной нейронной сети, что представляет собой значительную избыточность, а также позволяют только обнаруживать атаку.

Метод сертификационной защиты (Certified Defense) [81] предоставляет гарантию устойчивости модели к атакам на основе вредоносных возмущений определенной нормы. Повышение устойчивости модели достигается с помощью вычисления нижней границы L_p нормы возмущения для проведения успешной атаки и применения преобразований, направленных на нарушение целостности искажения [81, 82]. Однако такие методы защиты, как и состязательное обучение, негативно влияют на показатели качества модели. Кроме того, методы, реализующие сертификационную защиту, позволяют организовать защиту модели от возмущений до определенной нормы. Иными словами, указанный метод защиты не позволяет организовать защиту от множества различных атак.

Противодействие атакам на основе вредоносных возмущений может быть реализовано посредством шумоподавления [80, 83]. Такие методы защиты предполагают обнаружение и классификацию атаки с последующим устранением искажения. Однако методы шумоподавления не позволяют оказывать эффективное противодействие некоторым атакам, в том числе FGSM [18, 19] и BIM [23]. Кроме того, они предполагают сигнатурное определение атак, что не позволяет организовать защиту от неизвестных системе атак.

Достоинства и недостатки рассмотренных методов приведены в табл. 3.

Согласно табл. 3, существующие методы имеют значительные недостатки. Наиболее распространенный недостаток методов противодействия — использование дополнительной нейронной сети, что требует увеличения вычислительных мощностей. Этот факт накладывает значительные ограничения на область применения методов, обладающих указанным недостатком. Еще один немаловажный недостаток — невозможность устранения возмущения, что в свою очередь ограничивает область применения. Так, например, методы защиты, не позволяющие устранять возмущение неэффективны в системах реального времени, таких как беспилотный транспорт. Кроме того, для некоторых методов защиты имеет место снижение показателей качества целевой модели.

Возможные подходы к обнаружению и устранению вредоносных возмущений

Угрозы, связанные с атаками на основе вредоносных искажений — актуальны и критичны для многих систем, реализующих технологии искусственного интеллекта в различных прикладных областях. В то же время недостатки и ограничения существующих ме-

Таблица 3. Сравнение методов противодействия атакам на основе вредоносных возмущений
Table 3. Comparison of defense methods against attacks based on malicious perturbation

Метод	Достоинства	Недостатки
Состязательное обучение [17, 77, 78]	— позволяет игнорировать вредоносное возмущение; — не предполагает дополнительных вычислений	— снижение показателей качества целевой модели; — контроль репрезентативности обучающей выборки; — необходимость переобучения целевой нейронной сети
Защитная дистилляция [71]	— не оказывает значительного влияния на качество целевой нейронной сети; — позволяет обнаруживать атаки	— использование дополнительной нейронной сети; — неустойчивость к некоторым методам атаки; — не позволяет устранять возмущение
Сжатие параметров [79]	— не оказывает значительного влияния на качество целевой нейронной сети; — позволяет обнаруживать атаки	— использование дополнительной нейронной сети; — не позволяет устранять возмущение; — неприменим в системах, где качество изображения критически важно
Сертификационная защита [81, 82]	— позволяет устранять вредоносное возмущение; — сохранение качества при использовании для защиты нейронных сетей схожего функционала; — гарантия защиты от возмущения до определенной нормы	— снижение показателей качества целевой модели; — сложность адаптации ко множеству различных атак
Шумоподавление [80, 83]	— позволяет устранить вредоносное возмущение; — повышение качества изображения	— использование дополнительной нейронной сети; — зависимость эффективности от целевой модели; — неустойчивость к некоторым методам атаки; — низкая эффективность против неизвестных атак; — возможность удаления значимой информации.

тодов защиты не позволяют оказывать эффективного противодействия указанным атакам. Ввиду изложенного противоречия возникает необходимость улучшения существующих и разработки новых подходов к противодействию.

Обнаружение вредоносного возмущения возможно посредством статистического анализа шумовой компоненты изображения, в том числе при помощи дискретного косинусного преобразования и дискретного преобразования Фурье. Для анализа результатов преобразований возможно использование нечетного байесовского классификатора (Bayesian fuzzy clustering) [84], а для обработки шумовой компоненты в изначальном виде — критерия хи-квадрат (Chi-squared test) [85]. Данный подход может иметь ограничение по минимальной норме возмущения.

Устранение возмущения возможно при внесении обратимых визуальных модификаций и искажений изображения. Тогда использование прямых и обратных преобразований в различном порядке потенциально позволит нарушить целостность вредоносного возмущения. Заметим, что предложенный подход может иметь ограничение по максимальной норме устраняемого искажения и может быть неприменим в некоторых прикладных областях.

В то же время эффективное противодействие атакам, оптимизированным по псевдонорме L_0, предполагает использование иных подходов, описанных далее. Обнаружение вредоносного искажения, характерного для указанных атак, возможно посредством статистических методов, например: Z-оценки, гистограммного анализа, вычисления расстояния Махаланобиса [86]. Ввиду наличия определенных недостатков в указанных методах, возможно использование их комбинации. Кроме того, некоторые методы также вычисляют предполагаемые RGB-компоненты искаженной области.

Комбинация предложенных подходов потенциально позволяет обнаруживать и устранять вредоносные возмущения вне зависимости от их нормы. Следует отметить, что описанные подходы применимы исключительно для ограниченных искажений.

Заключение

Угрозы, связанные с атаками на основе вредоносных искажений, являются актуальными и критичными для многих систем, реализующих технологии искус-

ственного интеллекта в различных прикладных областях. В то же время недостатки и ограничения существующих методов защиты не позволяют оказывать эффективного противодействия указанным атакам. Из-за изложенного противоречия возникает необходимость улучшения существующих и разработки новых подходов к противодействию. В работе рассмотрены некоторые возможные варианты решения противоречия.

В дальнейшей работе предполагается реализация и проверка эффективности предложенных подходов, уточнение их ограничений и возможных областей применения. Кроме того, возможно формулирование подходов для противодействия неограниченным вредоносным возмущениям.

Литература

1. Goldberg Y. A primer on neural network models for natural language processing // Journal of Artificial Intelligence Research. 2016. V. 57. P. 345–420. https://doi.org/10.1613/jair.4992
2. Nassif A.B., Shahin I., Attili I., Azzeh M., Shaalan K. Speech recognition using deep neural networks: A systematic review // IEEE Access. 2019. V. 7. P. 19143–19165. https://doi.org/10.1109/access.2019.2896880
3. Almabdy S., Elrefaei L. Deep convolutional neural network-based approaches for face recognition // Applied Sciences. 2019. V. 9. N 20. P. 4397. https://doi.org/10.3390/app9204397
4. Khan M.Z., Harous S., Hassan S. U., Khan M. U. G., Iqbal R., Mumtaz S. Deep unified model for face recognition based on convolution neural network and edge computing // IEEE Access. 2019. V. 7. P. 72622–72633. https://doi.org/10.1109/access.2019.2918275
5. Zhang Y., Shi D., Zhan X., Cao D., Zhu K., Li Z. Slim-ResCNN: A deep residual convolutional neural network for fingerprint liveness detection // IEEE Access. 2019. V. 7. P. 91476–91487. https://doi.org/10.1109/access.2019.2927357
6. Sarvamangala D.R., Kulkarni R.V. Convolutional neural networks in medical image understanding: a survey // Evolutionary Intelligence. 2022. V. 15. N 1. P. 1–22. https://doi.org/10.1007/s12065-020-00540-3
7. Mahmood M., Al-Khateeb B., Alwash W. A review on neural networks approach on classifying cancers // IAES International Journal of Artificial Intelligence. 2020. V. 9. N 2. P. 317–326. http://doi.org/10.11591/ijai.v9.i2.pp317-326
8. Singh V., Singh S., Gupta P. Real-time anomaly recognition through CCTV using neural networks // Procedia Computer Science. 2020. V. 173. P. 254–263. https://doi.org/10.1016/j.procs.2020.06.030
9. Severino A., Curto S., Barberi S., Arena F., Pau G. Autonomous vehicles: an analysis both on their distinctiveness and the potential impact on urban transport systems // Applied Sciences. 2021. V. 11. N 8. P. 3604. https://doi.org/10.3390/app11083604
10. Wang L., Fan X., Chen J., Cheng J., Tan J., Ma X. 3D object detection based on sparse convolution neural network and feature fusion for autonomous driving in smart cities // Sustainable Cities and Society. 2020. V. 54. P. 102002. https://doi.org/10.1016/j.scs.2019.102002
11. Chen L., Lin S., Lu X., Cao D., Wu H., Guo C., Wang F. Y. Deep neural network based vehicle and pedestrian detection for autonomous driving: A survey // IEEE Transactions on Intelligent Transportation Systems. 2021. V. 22. N 6. P. 3234–3246. https://doi.org/10.1109/tits.2020.2993926
12. Chen P. Y., Liu S. Holistic adversarial robustness of deep learning models // Proceedings of the AAAI Conference on Artificial Intelligence. 2023. V. 37. N 13. P. 15411–15420. https://doi.org/10.1609/aaai.v37i13.26797
13. Huang X., Kroening D., Ruan W., Sharp J., Sun Y., Thamo E., Min W., Yi X. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability // Computer Science Review. 2020. V. 37. P. 100270. https://doi.org/10.1016/j.cosrev.2020.100270
14. Szegedy C., Zaremba W., Sutskever I., Bruna J., Erhan D., Goodfellow I., Fergus R. Intriguing properties of neural networks // arXiv. 2013. arXiv:1312.6199. https://doi.org/10.48550/arXiv.1312.6199
15. Song Y., Shu R., Kushman N., Ermon S. Constructing unrestricted adversarial examples with generative models // Advances in Neural Information Processing Systems. 2018. V. 31.
16. Sayghe A., Zhao J., Konstantinou C. Evasion attacks with adversarial deep learning against power system state estimation // Proc. of the 2020 IEEE Power & Energy Society General Meeting (PESGM). 2020. P. 1–5. https://doi.org/10.1109/pesgm41954.2020.9281719

References

1. Goldberg Y. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 2016, vol. 57, pp. 345–420. https://doi.org/10.1613/jair.4992
2. Nassif A.B., Shahin I., Attili I., Azzeh M., Shaalan K. Speech recognition using deep neural networks: A systematic review. *IEEE Access*, 2019, vol. 7, pp. 19143–19165. https://doi.org/10.1109/access.2019.2896880
3. Almabdy S., Elrefaei L. Deep convolutional neural network-based approaches for face recognition. *Applied Sciences*, 2019, vol. 9, no. 20, pp. 4397. https://doi.org/10.3390/app9204397
4. Khan M.Z., Harous S., Hassan S. U., Khan M. U. G., Iqbal R., Mumtaz S. Deep unified model for face recognition based on convolution neural network and edge computing. *IEEE Access*, 2019, vol. 7, pp. 72622–72633. https://doi.org/10.1109/access.2019.2918275
5. Zhang Y., Shi D., Zhan X., Cao D., Zhu K., Li Z. Slim-ResCNN: A deep residual convolutional neural network for fingerprint liveness detection. *IEEE Access*, 2019, vol. 7, pp. 91476–91487. https://doi.org/10.1109/access.2019.2927357
6. Sarvamangala D.R., Kulkarni R.V. Convolutional neural networks in medical image understanding: a survey. *Evolutionary Intelligence*, 2022, vol. 15, no. 1, pp. 1–22. https://doi.org/10.1007/s12065-020-00540-3
7. Mahmood M., Al-Khateeb B., Alwash W. A review on neural networks approach on classifying cancers. *IAES International Journal of Artificial Intelligence*, 2020, vol. 9, no. 2, pp. 317–326. http://doi.org/10.11591/ijai.v9.i2.pp317-326
8. Singh V., Singh S., Gupta P. Real-time anomaly recognition through CCTV using neural networks. *Procedia Computer Science*, 2020, vol. 173, pp. 254–263. https://doi.org/10.1016/j.procs.2020.06.030
9. Severino A., Curto S., Barberi S., Arena F., Pau G. Autonomous vehicles: an analysis both on their distinctiveness and the potential impact on urban transport systems. *Applied Sciences*, 2021, vol. 11, no. 8, pp. 3604. https://doi.org/10.3390/app11083604
10. Wang L., Fan X., Chen J., Cheng J., Tan J., Ma X. 3D object detection based on sparse convolution neural network and feature fusion for autonomous driving in smart cities. *Sustainable Cities and Society*, 2020, vol. 54, pp. 102002. https://doi.org/10.1016/j.scs.2019.102002
11. Chen L., Lin S., Lu X., Cao D., Wu H., Guo C., Wang F. Y. Deep neural network based vehicle and pedestrian detection for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 2021, vol. 22, no. 6, pp. 3234–3246. https://doi.org/10.1109/tits.2020.2993926
12. Chen P. Y., Liu S. Holistic adversarial robustness of deep learning models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, vol. 37, no. 13, pp. 15411–15420. https://doi.org/10.1609/aaai.v37i13.26797
13. Huang X., Kroening D., Ruan W., Sharp J., Sun Y., Thamo E., Min W., Yi X. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer Science Review*, 2020, vol. 37, pp. 100270. https://doi.org/10.1016/j.cosrev.2020.100270
14. Szegedy C., Zaremba W., Sutskever I., Bruna J., Erhan D., Goodfellow I., Fergus R. Intriguing properties of neural networks. *arXiv*, 2013, arXiv:1312.6199. https://doi.org/10.48550/arXiv.1312.6199
15. Song Y., Shu R., Kushman N., Ermon S. Constructing unrestricted adversarial examples with generative models. *Advances in Neural Information Processing Systems*, 2018, vol. 31.
16. Sayghe A., Zhao J., Konstantinou C. Evasion attacks with adversarial deep learning against power system state estimation. *Proc. of the 2020 IEEE Power & Energy Society General Meeting (PESGM)*, 2020, pp. 1–5. https://doi.org/10.1109/pesgm41954.2020.9281719

17. Goodfellow I.J., Shlens J., Szegedy C. Explaining and harnessing adversarial examples // arXiv. 2014. arXiv:1412.6572. https://doi.org/10.48550/arXiv.1412.6572
18. Paul R., Schabath M., Gillies R., Hall L., Goldgof D. Mitigating adversarial attacks on medical image understanding systems // Proc. of the 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). 2020. P. 1517–1521. https://doi.org/10.1109/isbi45749.2020.9098740
19. Rozsa A., Rudd E.M., Boult T.E. Adversarial diversity and hard positive generation // Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2016. P. 25–32. https://doi.org/10.1109/cvprw.2016.58
20. Dong Y., Liao F., Pang T., Su H., Zhu J., Hu X., Li J. Boosting adversarial attacks with momentum // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018. P. 9185–9193. https://doi.org/10.1109/cvpr.2018.00957
21. Miyato T., Maeda S.I., Koyama M., Ishii S. Virtual adversarial training: a regularization method for supervised and semi-supervised learning // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2019. V. 41. N 8. P. 1979–1993. https://doi.org/10.1109/tpami.2018.2858821
22. Kurakin A., Goodfellow I.J., Bengio S. Adversarial examples in the physical world // Artificial Intelligence Safety and Security. Chapman and Hall/CRC, 2018. P. 99–112. https://doi.org/10.1201/9781351251389-8
23. Mądry A., Makelov A., Schmidt L., Tsipras D., Vladu A. Towards deep learning models resistant to adversarial attacks // Stat. 2017. V. 1050. P. 9.
24. Xie C., Zhang Z., Zhou Y., Bai S., Wang J., Ren Z., Yuille A.L. Improving transferability of adversarial examples with input diversity // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019. P. 2730–2739. https://doi.org/10.1109/cvpr.2019.00284
25. Dong X., Han J., Chen D., Liu J., Bian H., Ma Z., Li H., Wang X., Zhang W., Yu N. Robust superpixel-guided attentional adversarial attack // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020. P. 12895–12904. https://doi.org/10.1109/cvpr42600.2020.01291
26. Sriramanan G., Addepalli S., Baburaj A. Guided adversarial attack for evaluating and enhancing adversarial defenses // Advances in Neural Information Processing Systems. 2020. V. 33. P. 20297–20308.
27. Rony J., Hafemann L.G., Oliveira L.S., Ayed I.B., Sabourin R., Granger E. Decoupling direction and norm for efficient gradient-based L2 adversarial attacks and defenses // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019. P. 4322–4330. https://doi.org/10.1109/cvpr.2019.00445
28. Moosavi-Dezfooli S.M., Fawzi A., Frossard P. DeepFool: a simple and accurate method to fool deep neural networks // Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016. P. 2574–2582. https://doi.org/10.1109/cvpr.2016.282
29. Carlini N., Wagner D. Towards evaluating the robustness of neural networks // Proc. of the IEEE Symposium on Security and Privacy (SP). 2017. P. 39–57. https://doi.org/10.1109/sp.2017.49
30. Yao Z., Gholami A., Xu P., Keutzer K., Mahoney M. W. Trust region based adversarial attack on neural networks // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019. P. 11350–11359. https://doi.org/10.1109/cvpr.2019.01161
31. Papernot N., McDaniel P., Jha S., Fredrikson M., Celik Z. B., Swami A. The limitations of deep learning in adversarial settings // Proc. of the 2016 IEEE European Symposium on Security and Privacy (EuroS&P). 2016. P. 372–387. https://doi.org/10.1109/eurosp.2016.36
32. Su J., Vargas D.V., Sakurai K. One pixel attack for fooling deep neural networks // IEEE Transactions on Evolutionary Computation. 2019. V. 23. N 5. P. 828–841. https://doi.org/10.1109/tevc.2019.2890858
33. Moosavi-Dezfooli S.M., Fawzi A., Fawzi O., Frossard P. Universal adversarial perturbations // Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017. P. 1765–1773. https://doi.org/10.1109/cvpr.2017.17
34. Brendel W., Rauber J., Bethge M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models // Advances in Reliably Evaluating and Improving Adversarial Robustness. 2021. P. 77.
35. Chen J., Jordan M.I., Wainwright M.J. HopSkipJumpAttack: A query-efficient decision-based attack // Proc. of the 2020 IEEE Symposium on Security and Privacy (SP). 2020. P. 1277–1294. https://doi.org/10.1109/sp40000.2020.00045

17. Goodfellow I.J., Shlens J., Szegedy C. Explaining and harnessing adversarial examples. *arXiv*, 2014, arXiv:1412.6572. https://doi.org/10.48550/arXiv.1412.6572
18. Paul R., Schabath M., Gillies R., Hall L., Goldgof D. Mitigating adversarial attacks on medical image understanding systems. *Proc. of the 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 2020, pp. 1517–1521. https://doi.org/10.1109/isbi45749.2020.9098740
19. Rozsa A., Rudd E.M., Boult T.E. Adversarial diversity and hard positive generation. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2016, pp. 25–32. https://doi.org/10.1109/cvprw.2016.58
20. Dong Y., Liao F., Pang T., Su H., Zhu J., Hu X., Li J. Boosting adversarial attacks with momentum. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9185–9193. https://doi.org/10.1109/cvpr.2018.00957
21. Miyato T., Maeda S.I., Koyama M., Ishii S. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, vol. 41, no. 8, pp. 1979–1993. https://doi.org/10.1109/tpami.2018.2858821
22. Kurakin A., Goodfellow I.J., Bengio S. Adversarial examples in the physical world. *Artificial Intelligence Safety and Security*. Chapman and Hall/CRC, 2018, pp. 99–112. https://doi.org/10.1201/9781351251389-8
23. Mądry A., Makelov A., Schmidt L., Tsipras D., Vladu A. Towards deep learning models resistant to adversarial attacks. *Stat*, 2017, vol. 1050, pp. 9.
24. Xie C., Zhang Z., Zhou Y., Bai S., Wang J., Ren Z., Yuille A.L. Improving transferability of adversarial examples with input diversity. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2730–2739. https://doi.org/10.1109/cvpr.2019.00284
25. Dong X., Han J., Chen D., Liu J., Bian H., Ma Z., Li H., Wang X., Zhang W., Yu N. Robust superpixel-guided attentional adversarial attack. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12895–12904. https://doi.org/10.1109/cvpr42600.2020.01291
26. Sriramanan G., Addepalli S., Baburaj A. Guided adversarial attack for evaluating and enhancing adversarial defenses. *Advances in Neural Information Processing Systems*, 2020, vol. 33, pp. 20297–20308.
27. Rony J., Hafemann L.G., Oliveira L.S., Ayed I.B., Sabourin R., Granger E. Decoupling direction and norm for efficient gradient-based L2 adversarial attacks and defenses. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4322–4330. https://doi.org/10.1109/cvpr.2019.00445
28. Moosavi-Dezfooli S.M., Fawzi A., Frossard P. DeepFool: a simple and accurate method to fool deep neural networks. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2574–2582. https://doi.org/10.1109/cvpr.2016.282
29. Carlini N., Wagner D. Towards evaluating the robustness of neural networks. *Proc. of the IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 39–57. https://doi.org/10.1109/sp.2017.49
30. Yao Z., Gholami A., Xu P., Keutzer K., Mahoney M. W. Trust region based adversarial attack on neural networks. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 11350–11359. https://doi.org/10.1109/cvpr.2019.01161
31. Papernot N., McDaniel P., Jha S., Fredrikson M., Celik Z. B., Swami A. The limitations of deep learning in adversarial settings. *Proc. of the 2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, 2016, pp. 372–387. https://doi.org/10.1109/eurosp.2016.36
32. Su J., Vargas D.V., Sakurai K. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 2019, vol. 23, no. 5, pp. 828–841. https://doi.org/10.1109/tevc.2019.2890858
33. Moosavi-Dezfooli S.M., Fawzi A., Fawzi O., Frossard P. Universal adversarial perturbations. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1765–1773. https://doi.org/10.1109/cvpr.2017.17
34. Brendel W., Rauber J., Bethge M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *Advances in Reliably Evaluating and Improving Adversarial Robustness*, 2021, pp. 77.
35. Chen J., Jordan M.I., Wainwright M.J. HopSkipJumpAttack: A query-efficient decision-based attack. *Proc. of the 2020 IEEE Symposium*

36. Liu Y., Moosavi-Dezfooli S.M., Frossard P. A geometry-inspired decision-based attack // Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV). 2019. P. 4890–4898. https://doi.org/10.1109/iccv.2019.00499
37. Rahmati A., Moosavi-Dezfooli, S.M., Frossard P., Dai H. GeoDA: a geometric framework for black-box adversarial attacks // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020. P. 8446–8455. https://doi.org/10.1109/cvpr42600.2020.00847
38. Du J., Zhang H., Zhou J.T., Yang Y., Feng J. Query-efficient meta attack to deep neural networks // Proc. of the International Conference on Learning Representations. 2020.
39. Li J., Ji R., Liu H., Liu J., Zhong B., Deng C., Tian Q. Projection & probability-driven black-box attack // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020. P. 362–371. https://doi.org/10.1109/cvpr42600.2020.00044
40. Li H., Xu X., Zhang X., Yang S., Li B. QEBA: Query-efficient boundary-based blackbox attack // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020. P. 1221–1230. https://doi.org/10.1109/cvpr42600.2020.00130
41. Cheng M., Singh S., Chen P., Chen P.Y., Liu S., Hsieh C.J. Sign-OPT: A query-efficient hard-label adversarial attack // Proc. of the International Conference on Learning Representations. 2020.
42. Brunner T., Diehl F., Le M.T., Knoll A. Guessing smart: Biased sampling for efficient black-box adversarial attacks // Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV). 2019. P. 4958–4966. https://doi.org/10.1109/iccv.2019.00506
43. Maho T., Furon T., Le Merrer E. SurFree: a fast surrogate-free black-box attack // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021. P. 10430–10439. https://doi.org/10.1109/cvpr46437.2021.01029
44. Shi Y., Han Y., Tian Q. Polishing decision-based adversarial noise with a customized sampling // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020. P. 1030–1038. https://doi.org/10.1109/cvpr42600.2020.00111
45. Huang Z., Zhang T. Black-box adversarial attack with transferable model-based embedding // Proc. of the International Conference on Learning Representations. 2020.
46. Zhou M., Wu J., Liu Y., Liu S., Zhu C. DaST: Data-free substitute training for adversarial attacks // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020. P. 234–243. https://doi.org/10.1109/cvpr42600.2020.00031
47. Zou J., Pan Z., Qiu J., Liu X., Rui T., Li W. Improving the transferability of adversarial examples with resized-diverse-inputs, diversity-ensemble and region fitting // Lecture Notes in Computer Science. 2020. V. 12367. P. 563–579. https://doi.org/10.1007/978-3-030-58542-6_34
48. Wang X., He K. Enhancing the transferability of adversarial attacks through variance tuning // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021. P. 1924–1933. https://doi.org/10.1109/cvpr46437.2021.00196
49. Wu W., Su Y., Lyu M.R., King I. Improving the transferability of adversarial samples with adversarial transformations // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021. P. 9024–9033. https://doi.org/10.1109/cvpr46437.2021.00891
50. Hosseini H., Poovendran R. Semantic adversarial examples // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2018. P. 1614–1619. https://doi.org/10.1109/cvprw.2018.00212
51. Engstrom L., Tran B., Tsipras D., Schmidt L., Madry A. A rotation and a translation suffice: Fooling cnns with simple transformations [Электронный ресурс]. URL: https://openreview.net/forum?id=BJfvknCqFQ (дата обращения: 29.05.2023).
52. Joshi A., Mukherjee A., Sarkar S., Hegde C. Semantic adversarial attacks: Parametric transformations that fool deep classifiers // Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV). 2019. P. 4773–4783. https://doi.org/10.1109/iccv.2019.00487
53. Liu A., Wang J., Liu X., Cao B., Zhang C., Yu H. Bias-based universal adversarial patch attack for automatic check-out // Lecture Notes in Computer Science. 2020. V. 12358. P. 395–410. https://doi.org/10.1007/978-3-030-58601-0_24
54. Swathi P., Sk S. DeepFake creation and detection: A survey // Proc. of the 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA). 2021. P. 584–588. https://doi.org/10.1109/icirca51532.2021.9544522

on Security and Privacy (SP), 2020, pp. 1277–1294. https://doi.org/10.1109/sp40000.2020.00045
36. Liu Y., Moosavi-Dezfooli S.M., Frossard P. A geometry-inspired decision-based attack. *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 4890–4898. https://doi.org/10.1109/iccv.2019.00499
37. Rahmati A., Moosavi-Dezfooli, S.M., Frossard P., Dai H. GeoDA: a geometric framework for black-box adversarial attacks. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8446–8455. https://doi.org/10.1109/cvpr42600.2020.00847
38. Du J., Zhang H., Zhou J.T., Yang Y., Feng J. Query-efficient meta attack to deep neural networks. *Proc. of the International Conference on Learning Representations*, 2020.
39. Li J., Ji R., Liu H., Liu J., Zhong B., Deng C., Tian Q. Projection & probability-driven black-box attack. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 362–371. https://doi.org/10.1109/cvpr42600.2020.00044
40. Li H., Xu X., Zhang X., Yang S., Li B. QEBA: Query-efficient boundary-based blackbox attack. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1221–1230. https://doi.org/10.1109/cvpr42600.2020.00130
41. Cheng M., Singh S., Chen P., Chen P.Y., Liu S., Hsieh C.J. Sign-OPT: A query-efficient hard-label adversarial attack. *Proc. of the International Conference on Learning Representations*, 2020.
42. Brunner T., Diehl F., Le M.T., Knoll A. Guessing smart: Biased sampling for efficient black-box adversarial attacks. *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 4958–4966. https://doi.org/10.1109/iccv.2019.00506
43. Maho T., Furon T., Le Merrer E. SurFree: a fast surrogate-free black-box attack. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 10430–10439. https://doi.org/10.1109/cvpr46437.2021.01029
44. Shi Y., Han Y., Tian Q. Polishing decision-based adversarial noise with a customized sampling. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1030–1038. https://doi.org/10.1109/cvpr42600.2020.00111
45. Huang Z., Zhang T. Black-box adversarial attack with transferable model-based embedding. *Proc. of the International Conference on Learning Representations*, 2020.
46. Zhou M., Wu J., Liu Y., Liu S., Zhu C. DaST: Data-free substitute training for adversarial attacks. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 234–243. https://doi.org/10.1109/cvpr42600.2020.00031
47. Zou J., Pan Z., Qiu J., Liu X., Rui T., Li W. Improving the transferability of adversarial examples with resized-diverse-inputs, diversity-ensemble and region fitting. *Lecture Notes in Computer Science*, 2020, vol. 12367, pp. 563–579. https://doi.org/10.1007/978-3-030-58542-6_34
48. Wang X., He K. Enhancing the transferability of adversarial attacks through variance tuning. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 1924–1933. https://doi.org/10.1109/cvpr46437.2021.00196
49. Wu W., Su Y., Lyu M.R., King I. Improving the transferability of adversarial samples with adversarial transformations. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 9024–9033. https://doi.org/10.1109/cvpr46437.2021.00891
50. Hosseini H., Poovendran R. Semantic adversarial examples. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 1614–1619. https://doi.org/10.1109/cvprw.2018.00212
51. Engstrom L., Tran B., Tsipras D., Schmidt L., Madry A. *A rotation and a translation suffice: Fooling cnns with simple transformations*. Available at: https://openreview.net/forum?id=BJfvknCqFQ (accessed: 29.05.2023).
52. Joshi A., Mukherjee A., Sarkar S., Hegde C. Semantic adversarial attacks: Parametric transformations that fool deep classifiers. *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 4773–4783. https://doi.org/10.1109/iccv.2019.00487
53. Liu A., Wang J., Liu X., Cao B., Zhang C., Yu H. Bias-based universal adversarial patch attack for automatic check-out. *Lecture Notes in Computer Science*, 2020, vol. 12358, pp. 395–410. https://doi.org/10.1007/978-3-030-58601-0_24
54. Swathi P., Sk S. DeepFake creation and detection: A survey. *Proc. of the 2021 Third International Conference on Inventive Research in*

55. Chadha A., Kumar V., Kashyap S., Gupta M. Deepfake: An Overview // Lecture Notes in Networks and Systems. 2021. V. 203. P. 557–566. https://doi.org/10.1007/978-981-16-0733-2_39
56. Nakka K.K., Salzmann M. Indirect local attacks for context-aware semantic segmentation networks // Lecture Notes in Computer Science. 2020. V. 12350. P. 611–628. https://doi.org/10.1007/978-3-030-58558-7_36
57. He Y., Rahimian S., Schiele B., Fritz M. Segmentations-leak: Membership inference attacks and defenses in semantic image segmentation // Lecture Notes in Computer Science. 2020. V. 12368. P. 519–535. https://doi.org/10.1007/978-3-030-58592-1_31
58. Choi J.H. Zhang H., Kim J.H., Hsieh C.J., Lee J.S. Evaluating robustness of deep image super-resolution against adversarial attacks // Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV). 2019. P. 303–311. https://doi.org/10.1109/iccv.2019.00039
59. Jiang L., Ma X., Chen S., Bailey J., Jiang Y.G. Black-box adversarial attacks on video recognition models // Proc. of the 27th ACM International Conference on Multimedia. 2019. P. 864–872. https://doi.org/10.1145/3343031.3351088
60. Li S., Aich A., Zhu S., Asif S., Song C., Roy-Chowdhury A., Krishnamurthy S. Adversarial attacks on black box video classifiers: Leveraging the power of geometric transformations // Advances in Neural Information Processing Systems. 2021. V. 34. P. 2085–2096.
61. Chen X., Li S., Huang H. Adversarial attack and defense on deep neural network-based voice processing systems: An overview // Applied Sciences. 2021. V. 11. N 18. P. 8450. https://doi.org/10.3390/app11188450
62. Kwon H., Kim Y., Yoon H., Choi D. Selective audio adversarial example in evasion attack on speech recognition system // IEEE Transactions on Information Forensics and Security. 2020. V. 15. P. 526–538. https://doi.org/10.1109/tifs.2019.2925452
63. Usama M., Qayyum A., Qadir J., Al-Fuqaha A. Black-box adversarial machine learning attack on network traffic classification // Proc. of the 15th International Wireless Communications & Mobile Computing Conference (IWCMC). 2019. P. 84–89. https://doi.org/10.1109/iwcmc.2019.8766505
64. Imam N.H., Vassilakis V.G. A survey of attacks against twitter spam detectors in an adversarial environment // Robotics. 2019. V. 8. N 3. P. 50. https://doi.org/10.3390/robotics8030050
65. Zhong H., Liao C., Squicciarini A.C., Zhu S., Miller D. Backdoor embedding in convolutional neural network models via invisible perturbation // Proc. of the Tenth ACM Conference on Data and Application Security and Privacy. 2020. P. 97–108. https://doi.org/10.1145/3374664.3375751
66. Liu X., Yang H., Liu Z., Song L., Li H., Chen Y. Dpatch: An adversarial patch attack on object detectors // arXiv. 2018. arXiv:1806.02299. https://doi.org/10.48550/arXiv.1806.02299
67. Liu Y., Ma X., Bailey J., Lu F. Reflection backdoor: A natural backdoor attack on deep neural networks // Lecture Notes in Computer Science. 2020. V. 12355. P. 182–199. https://doi.org/10.1007/978-3-030-58607-2_11
68. Nguyen A., Tran A. WaNet - imperceptible warping-based backdoor attack // Proc. of the International Conference on Learning Representations. 2021.
69. Костюмов В.В. Обзор и систематизация атак уклонением на модели компьютерного зрения // International Journal of Open Information Technologies. 2022. Т. 10. № 10. С. 11–20.
70. Papernot N., McDaniel P., Wu X., Jha S., Swami A. Distillation as a defense to adversarial perturbations against deep neural networks // Proc. of the 2016 IEEE Symposium on Security and Privacy (SP). 2016. P. 582–597. https://doi.org/10.1109/sp.2016.41
71. Steihaug T. The conjugate gradient method and trust regions in large scale optimization // SIAM Journal on Numerical Analysis. 1983. V. 20. N 3. P. 626–637. https://doi.org/10.1137/0720042
72. Curtis A.R., Powell M.J.D., Reid J.K. On the estimation of sparse Jacobian matrices // IMA Journal of Applied Mathematics. 1974. V. 13. N 1. P. 117–120. https://doi.org/10.1093/imamat/13.1.117
73. Niebur E. Saliency map // Scholarpedia. 2007. V. 2. N 8. C. 2675. https://doi.org/10.4249/scholarpedia.2675
74. Das S., Suganthan P.N. Differential evolution: A survey of the state-of-the-art // IEEE Transactions on Evolutionary Computation. 2011. V. 15. N 1. P. 4–31. https://doi.org/10.1109/tevc.2010.2059031
75. Badrinarayanan V., Kendall A., Cipolla R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation // IEEE Transactions on Pattern Analysis and Machine Intelligence.

Computing Applications (ICIRCA), 2021, pp. 584–588. https://doi.org/10.1109/icirca51532.2021.9544522
55. Chadha A., Kumar V., Kashyap S., Gupta M. Deepfake: An Overview. *Lecture Notes in Networks and Systems*, 2021, vol. 203, pp. 557–566. https://doi.org/10.1007/978-981-16-0733-2_39
56. Nakka K.K., Salzmann M. Indirect local attacks for context-aware semantic segmentation networks. *Lecture Notes in Computer Science*, 2020, vol. 12350, pp. 611–628. https://doi.org/10.1007/978-3-030-58558-7_36
57. He Y., Rahimian S., Schiele B., Fritz M. Segmentations-leak: Membership inference attacks and defenses in semantic image segmentation. *Lecture Notes in Computer Science*, 2020, vol. 12368, pp. 519–535. https://doi.org/10.1007/978-3-030-58592-1_31
58. Choi J.H. Zhang H., Kim J.H., Hsieh C.J., Lee J.S. Evaluating robustness of deep image super-resolution against adversarial attacks. *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 303–311. https://doi.org/10.1109/iccv.2019.00039
59. Jiang L., Ma X., Chen S., Bailey J., Jiang Y.G. Black-box adversarial attacks on video recognition models. *Proc. of the 27th ACM International Conference on Multimedia*, 2019, pp. 864–872. https://doi.org/10.1145/3343031.3351088
60. Li S., Aich A., Zhu S., Asif S., Song C., Roy-Chowdhury A., Krishnamurthy S. Adversarial attacks on black box video classifiers: Leveraging the power of geometric transformations. *Advances in Neural Information Processing Systems*, 2021, vol. 34, pp. 2085–2096.
61. Chen X., Li S., Huang H. Adversarial attack and defense on deep neural network-based voice processing systems: An overview. *Applied Sciences*, 2021, vol. 11, no. 18, pp. 8450. https://doi.org/10.3390/app11188450
62. Kwon H., Kim Y., Yoon H., Choi D. Selective audio adversarial example in evasion attack on speech recognition system. *IEEE Transactions on Information Forensics and Security*, 2020, vol. 15, pp. 526–538. https://doi.org/10.1109/tifs.2019.2925452
63. Usama M., Qayyum A., Qadir J., Al-Fuqaha A. Black-box adversarial machine learning attack on network traffic classification. *Proc. of the 15th International Wireless Communications & Mobile Computing Conference (IWCMC)*, 2019, pp. 84–89. https://doi.org/10.1109/iwcmc.2019.8766505
64. Imam N.H., Vassilakis V.G. A survey of attacks against twitter spam detectors in an adversarial environment. *Robotics*, 2019, vol. 8, no. 3, pp. 50. https://doi.org/10.3390/robotics8030050
65. Zhong H., Liao C., Squicciarini A.C., Zhu S., Miller D. Backdoor embedding in convolutional neural network models via invisible perturbation. *Proc. of the Tenth ACM Conference on Data and Application Security and Privacy*, 2020, pp. 97–108. https://doi.org/10.1145/3374664.3375751
66. Liu X., Yang H., Liu Z., Song L., Li H., Chen Y. Dpatch: An adversarial patch attack on object detectors. *arXiv*, 2018, arXiv:1806.02299. https://doi.org/10.48550/arXiv.1806.02299
67. Liu Y., Ma X., Bailey J., Lu F. Reflection backdoor: A natural backdoor attack on deep neural networks. *Lecture Notes in Computer Science*, 2020, vol. 12355, pp. 182–199. https://doi.org/10.1007/978-3-030-58607-2_11
68. Nguyen A., Tran A. WaNet - imperceptible warping-based backdoor attack. *Proc. of the International Conference on Learning Representations*, 2021.
69. Kostyumov V. A Survey and systematization of evasion attacks in computer vision. *International Journal of Open Information Technologies*, 2022, vol. 10, no. 10, pp. 11–20. (in Russian)
70. Papernot N., McDaniel P., Wu X., Jha S., Swami A. Distillation as a defense to adversarial perturbations against deep neural networks. *Proc. of the 2016 IEEE Symposium on Security and Privacy (SP)*, 2016, pp. 582–597. https://doi.org/10.1109/sp.2016.41
71. Steihaug T. The conjugate gradient method and trust regions in large scale optimization. *SIAM Journal on Numerical Analysis*, 1983, vol. 20, no. 3, pp. 626–637. https://doi.org/10.1137/0720042
72. Curtis A.R., Powell M.J.D., Reid J.K. On the estimation of sparse Jacobian matrices. *IMA Journal of Applied Mathematics*, 1974, vol. 13, no. 1, pp. 117–120. https://doi.org/10.1093/imamat/13.1.117
73. Niebur E. Saliency map. *Scholarpedia*, 2007, vol. 2, no. 8, pp. 2675. https://doi.org/10.4249/scholarpedia.2675
74. Das S., Suganthan P.N. Differential evolution: A survey of the state-of-the-art. *IEEE Transactions on Evolutionary Computation*, 2011, vol. 15, no. 1, pp. 4–31. https://doi.org/10.1109/tevc.2010.2059031

2017. V. 39. N 12. P. 2481–2495. https://doi.org/10.1109/tpami.2016.2644615
76. Lowd D., Meek C. Adversarial learning // Proc. of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. 2005. P. 641–647. https://doi.org/10.1145/1081870.1081950
77. Xu W., Evans D., Qi Y. Feature squeezing: Detecting adversarial examples in deep neural networks // Proc. of the 2018 Network and Distributed System Security Symposium (NDSS). 2018.
78. Liao F., Liang M., Dong Y., Pang T., Hu X., Zhu J. Defense against adversarial attacks using high-level representation guided denoiser // Proc. of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018. P. 1778–1787. https://doi.org/10.1109/cvpr.2018.00191
79. Zhang D., Ye M., Gong C., Zhu Z., Liu Q. Black-box certification with randomized smoothing: A functional optimization based framework // Advances in Neural Information Processing Systems. 2020. V. 33. P. 2316–2326.
80. Fischer M., Baader M., Vechev M. Certified defense to image transformations via randomized smoothing // Advances in Neural Information Processing Systems. 2020. V. 33. P. 8404–8417.
81. Yang R., Chen X.Q., Cao T.J. APE-GAN++: An improved APE-GAN to eliminate adversarial perturbations // IAENG International Journal of Computer Science. 2021. V. 48. N 3. P. 827–844.
82. Glenn T.C., Zare A., Gader P.D. Bayesian fuzzy clustering // IEEE Transactions on Fuzzy Systems. 2015. V. 23. N 5. P. 1545–1561. https://doi.org/10.1109/tfuzz.2014.2370676
83. Plackett R.L. Karl Pearson and the chi-squared test // International Statistical Review / Revue Internationale de Statistique. 1983. V. 51. N 1. P. 59–72. https://doi.org/10.2307/1402731
84. McLachlan G.J. Mahalanobis distance // Resonance. 1999. V. 4. N 6. P. 20–26. https://doi.org/10.1007/BF02834632

75. Badrinarayanan V., Kendall A., Cipolla R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, vol. 39, no. 12, pp. 2481–2495. https://doi.org/10.1109/tpami.2016.2644615
76. Lowd D., Meek C. Adversarial learning. *Proc. of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, 2005, pp. 641–647. https://doi.org/10.1145/1081870.1081950
77. Xu W., Evans D., Qi Y. Feature squeezing: Detecting adversarial examples in deep neural networks. *Proc. of the 2018 Network and Distributed System Security Symposium (NDSS)*, 2018.
78. Liao F., Liang M., Dong Y., Pang T., Hu X., Zhu J. Defense against adversarial attacks using high-level representation guided denoiser. *Proc. of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1778–1787. https://doi.org/10.1109/cvpr.2018.00191
79. Zhang D., Ye M., Gong C., Zhu Z., Liu Q. Black-box certification with randomized smoothing: A functional optimization based framework. *Advances in Neural Information Processing Systems*, 2020, vol. 33, pp. 2316–2326.
80. Fischer M., Baader M., Vechev M. Certified defense to image transformations via randomized smoothing. *Advances in Neural Information Processing Systems*, 2020, vol. 33, pp. 8404–8417.
81. Yang R., Chen X.Q., Cao T.J. APE-GAN++: An improved APE-GAN to eliminate adversarial perturbations. *IAENG International Journal of Computer Science*, 2021, vol. 48, no. 3, pp. 827–844.
82. Glenn T.C., Zare A., Gader P.D. Bayesian fuzzy clustering. *IEEE Transactions on Fuzzy Systems*, 2015, vol. 23, no. 5, pp. 1545–1561. https://doi.org/10.1109/tfuzz.2014.2370676
83. Plackett R.L. Karl Pearson and the chi-squared test. *International Statistical Review / Revue Internationale de Statistique*, 1983, vol. 51, no. 1, pp. 59–72. https://doi.org/10.2307/1402731
84. McLachlan G.J. Mahalanobis distance. *Resonance*, 1999, vol. 4, no. 6, pp. 20–26. https://doi.org/10.1007/BF02834632

Авторы

Есипов Дмитрий Андреевич — инженер, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0000-0003-4467-5117, some1else.d.ma@gmail.com
Бучаев Абдулхамид Яхьяевич — инженер, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57219568840, https://orcid.org/0009-0001-1058-9125, abdulhamid0055@yandex.ru
Керимбай Акылжан — инженер, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0009-0009-9945-9906, akerimbai@itmo.ru
Пузикова Яна Владиславовна — инженер, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0009-0007-7604-3022, Yanapuzikova19@ya.ru
Сайдумаров Семен Кириллович — студент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0009-0008-0774-9803, semen.say@ya.ru
Сулименко Никита Сергеевич — студент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0009-0007-3218-9249, n.s.sulimenko@mail.ru
Попов Илья Юрьевич — кандидат технических наук, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57202195632, https://orcid.org/0000-0002-6407-7934, ilyapopov27@gmail.com
Кармановский Николай Сергеевич — кандидат технических наук, доцент, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57192385103, https://orcid.org/0000-0002-0533-9893, karmanov50@mail.ru

Authors

Dmitry A. Esipov — Engineer, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0000-0003-4467-5117, some1else.d.ma@gmail.com
Abdulhamid Y. Buchaev — Engineer, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57219568840, https://orcid.org/0009-0001-1058-9125, abdulhamid0055@yandex.ru
Akylzhan Kerimbay — Engineer, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0009-0009-9945-9906, akerimbai@itmo.ru
Yana V. Puzikova — Engineer, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0009-0007-7604-3022, Yanapuzikova19@ya.ru
Semen K. Saidumarov — Student, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0009-0008-0774-9803, semen.say@ya.ru
Nikita S. Sulimenko — Student, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0009-0007-3218-9249, n.s.sulimenko@mail.ru
Ilya Yu. Popov — PhD, Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57202195632, https://orcid.org/0000-0002-6407-7934, ilyapopov27@gmail.com

Nikolay S. Karmanovskiy — PhD, Associate Professor, Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57192385103, https://orcid.org/0000-0002-0533-9893, karmanov50@mail.ru

Статья поступила в редакцию 27.03.2023
Одобрена после рецензирования 08.06.2023
Принята к печати 30.07.2023

Received 27.03.2023
Approved after reviewing 08.06.2023
Accepted 30.07.2023