

УДК 004.838.2

doi: 10.15622/rcai.2025.073

ИССЛЕДОВАНИЕ ВОЗМОЖНОСТЕЙ КОМБИНИРОВАНИЯ МЕТОДОВ ЗАЩИТЫ СИСТЕМ РАСПОЗНАВАНИЯ ИЗОБРАЖЕНИЙ ОТ ГРАДИЕНТНЫХ СОСТЯЗАТЕЛЬНЫХ АТАК¹

И.В. Котенко (*ivkote@comsec.spb.ru*)

И.Б. Саенко (*ibsaen@comsec.spb.ru*)

В.Е. Садовников (*bladimir1998@mail.ru*)

Санкт-Петербургский Федеральный исследовательский центр РАН,
Санкт-Петербург

В статье рассматривается подход к защите систем распознавания изображений от градиентных состязательных атак, основанный на комбинировании различных методов защиты, включающих зашумление, сжатие и нейросетевую очистку изображений. Комбинирование методов защиты предполагает нахождение оптимальных параметров, характеризующих эти методы, при которых точность распознавания становится максимальной. Оценка эффективности рассматриваемых методов защиты производится на датасете STL-10. Выбор этого датасета обусловлен его широким применением в задачах классификации изображений. Результаты экспериментов показывают, что комбинирование указанных методов защиты позволяет достигнуть максимальной точности распознавания изображений в условиях воздействия на них градиентных состязательных атак.

Ключевые слова: защита от атак, искусственный интеллект, состязательная атака, распознавание изображений, машинное обучение, зашумление, кибербезопасность.

Введение

Современные системы распознавания изображений на базе глубоких нейронных сетей демонстрируют высокую точность в задачах классификации, сегментации, детектирования объектов и других областях компью-

¹ Работа выполнена при частичной финансовой поддержке бюджетной темы FFZF-2025-0016.

терного зрения. Эти достижения широко применяются в таких критически важных сферах, как медицинская диагностика, системы видеонаблюдения, автономное вождение, биометрическая идентификация и интеллектуальные промышленные системы [Понкин и др., 2024]. Однако несмотря на достигнутый прогресс нейросетевые модели остаются уязвимыми к состязательным атакам – специально сконструированным входным данным (примерам), способным вызвать ошибочное поведение модели при визуальной неотличимости искажений для человека [Легашев и др., 2024].

Исследования показали, что даже минимальные возмущения, добавленные к исходному изображению, могут привести к резкому снижению точности классификации и полному нарушению работы модели [Баев и др., 2024]. Наиболее известным примером таких атак является метод Fast Gradient Sign Method (FGSM). При этом методе для создания возмущения, искажающего классификацию, используется однократное вычисление градиента функции потерь. Развитие этого подхода привело к появлению метода Iterative FGSM (IFGSM), в котором атака осуществляется через серию последовательных шагов, повышая эффективность вмешательства. В условиях ограниченного доступа к внутренним параметрам модели применяются методы «черного ящика», такие как Zeroth Order Optimization (ZOO), основанные на численной оценке градиента без использования внутренней информации о сети. Эти методы демонстрируют высокую эффективность при обходе защиты, даже при минимальной информации о целевой модели. Все градиентные атаки наглядно демонстрируют уязвимость глубоких моделей и подчеркивают необходимость разработки надежных средств защиты, особенно в контексте практического применения вне контролируемой среды.

На фоне растущей актуальности проблемы в последние годы активно исследуются подходы к повышению устойчивости моделей к подобным воздействиям. Среди них – модификации архитектур, обогащение обучающих выборок, использование робастных функций потерь, а также методы предобработки входных данных, направленные на нейтрализацию состязательных искажений. К числу последних относятся такие подходы, как JPEG-сжатие, добавление случайных шумов, а также нейросетевая очистка изображений с помощью автоэнкодеров.

Вклад и новизна статьи заключаются в следующем: (1) предложен новый подход к защите от градиентных состязательных атак, основанный на комбинировании методов защиты; (2) проведено экспериментальное исследование эффективности защиты в одиночном, попарном и комбинированном режимах применения методов; (3) определены оптимальные параметры методов защиты, при которых точность обнаружения атак становится максимальной. Эксперименты подтвердили, что комбинирование методов защиты с оптимальными параметрами дает наивысшую точность обнаружения градиентных атак.

1. Анализ текущего состояния исследований

Одним из наиболее эффективных направлений обеспечения устойчивости систем компьютерного зрения являются гибридные методы, сочетающие зашумление, сжатие и нейронную очистку входных изображений, позволяющие снижать влияние атак без большого ущерба для эффективности распознавания [Фомичева и др., 2023].

В [Li et al., 2023] описан метод ComDefend, основанный на адаптивном сжатии и восстановлении структуры изображения без изменения архитектуры классификатора. Он демонстрирует высокую устойчивость к атакам типа FGSM. Метод Compress and Restore [Naveen et al., 2023] использует итеративное JPEG-сжатие с переменным качеством и генеративное восстановление (AR-GAN). В работе [Gardella et al., 2022] подчеркивается эффективность сочетания шума и JPEG-сжатия для судебно-медицинского анализа изображений и восстановления изображений.

Гибридный метод HAT (Hybrid Adversarial Training) [Ali et al., 2022] дополняет обучение состязательными примерами, созданными путем объединения атак DeepFool и FGSM, тем самым увеличивая устойчивость моделей глубокого обучения в течение установленного периода времени против различных атак.

В [Jain, 2024] исследуется влияние различных типов атак, включая атаку FGSM, на точность алгоритма обнаружения и дорожных знаков YOLOv5. Результаты показывают, что алгоритм подвержен этим атакам, причем показатели неправильной классификации увеличиваются по мере увеличения величины возмущений.

В [Kotlyarov et al., 2023] рассмотрен алгоритм создания нейросети, ориентированной на распознавание образов, и рассмотрены несколько видов состязательных атак, включая FGSM. Получено подтверждение гипотезы о снижении точности распознавания нейросети при реализации состязательной атаки злоумышленником.

В [Tian et al., 2024] предложен подход к построению моделей глубокого обучения, устойчивых к воздействию различных состязательных атак, основанный на преобразовании изображений для удаления состязательных шумов. Экспериментальные результаты подтвердили потенциал методов преобразования изображений как надежной защиты от состязательных атак в системах классификации изображений на основе глубокого обучения, особенно в сочетании с передовыми архитектурами нейронных сетей, такими как ResNet50 и DenseNet121.

В [Khamaiseh et al., 2022] рассматриваются методы состязательных атак, включая FGSM, IFGSM и ZOO, с акцентом на объяснение математических концепций и терминологии, а также механизмы защиты с обсуж-

дением их эффективность в защите глубоких нейронных сетей от состязательных атак. Показано, что идея использовать нейронную очистку с автоэнкодерами, зашумление и сжатие изображений являются перспективными методами защиты от современных состязательных атак.

Авторами настоящей работы также изучался данный вопрос. Так, в [Kotenko et al., 2024] была предложена схема JPEG-сжатия и нейронной фильтрации, показавшая высокую эффективность на двух наборах данных. В [Kotenko и др., 2025] был успешно протестирован подход на основе JPEG и Neural Cleanse, позволивший почти полностью восстановить точность классификации на атакованных изображениях, особенно при использовании ансамблевых моделей.

Таким образом, анализ современных исследований в области защиты от градиентных состязательных атак показывает, что гибридные методы являются здесь одним из наиболее перспективных направлений. Настоящая работа посвящена исследованию комбинирования трех методов защиты.

2. Градиентные состязательные атаки

Метод FGSM стал одной из первых и наиболее широко используемых реализаций градиентных состязательных атак на нейросетевые классификаторы [Goodfellow et al., 2015]. Основная идея FGSM заключается в том, чтобы модифицировать входное изображение таким образом, чтобы оно вызывало ошибочную классификацию модели, при этом изменение оставалось незаметным для человеческого глаза.

Атака основывается на градиенте функции потерь по отношению к входным данным. Состязательное возмущение вычисляется как

$$\eta = \epsilon \operatorname{sign} \left(\nabla_x J(\theta, x, y) \right), \quad (2.1)$$

где η – добавляемое к изображению возмущение; ϵ – малая константа, определяющая интенсивность искажений; $J(\theta, x, y)$ – функция потерь модели; $\nabla_x J(\theta, x, y)$ – градиент функции потерь по отношению к входу. Метод предполагает однократное вычисление градиента функции потерь и направляет изменение пикселей входного изображения в сторону, в которой ошибка модели возрастает максимально быстро. Это делает FGSM не только эффективным, но и вычислительно дешевым методом, так как он не требует итеративного процесса оптимизации.

Метод IGSM представляет собой итеративное расширение метода FGSM, обеспечивающее более точную настройку состязательных возмущений [Jiakai, 2018]. Идея IGSM состоит в итеративном применении атаки FGSM к изображению с небольшим шагом, позволяя более точно направлять изображение в сторону неверной классификации, сохраняя при этом низкую заметность искажений.

Состязательное возмущение вычисляется по итеративной формуле

$$x^{(t+1)} = \text{Clip}_{x,\varepsilon} \left(x^{(t)} + \varepsilon \cdot \text{sign} \left(\nabla_x J(\theta, x, y) \right) \right), x^{(0)} = x, \quad (2.2)$$

где $x^{(t)}$ – изображение на итерации t ; ε – малый шаг изменения пикселей на каждой итерации; $J(\theta, x, y)$ – функция потерь; $\nabla_x J(\theta, x, y)$ – градиент функции потерь, $\text{Clip}_{x,\varepsilon}$ – операция отсечения, гарантирующая, что возмущенное изображение x останется в заданной области ε допустимых значений пикселей. За счет итеративного процесса IGSM демонстрирует более высокую эффективность по сравнению с одношаговым FGSM, однако требует больших вычислительных затрат.

Атака ZOO представляет собой метод «черного ящика», не требующий доступа к внутренним параметрам модели или ее градиентам [Chen et al., 2017]. В отличие от традиционных атак «белого ящика», она позволяет выполнять состязательные воздействия на нейросеть, имея доступ только к значениям выходов модели. Основная идея атаки состоит в численной аппроксимации градиентов функции потерь относительно входного изображения. Так как градиенты напрямую недоступны, ZOO использует метод конечных разностей, чтобы оценить производную по каждой координате (пикселю) входа:

$$\nabla_{x,i} f(x) \approx \frac{f(x + \varepsilon e_i) - f(x - \varepsilon e_i)}{2\varepsilon}, \quad (2.3)$$

где ∇ – аппроксимированная производная функция потерь; x – входное изображение; e_i – единичный вектор с единицей на i -й позиции и нулями в остальных координатах; $f(x)$ – функция потерь, зависящая от выхода модели и целевого класса. Полученные градиенты используются для итеративного обновления входного изображения. Вместо одновременного обновления всех координат, ZOO применяет стохастический координатный спуск – на каждой итерации обновляется одна или несколько координат, что позволяет существенно сократить количество вызовов модели. ZOO показал эффективность, сравнимую с «белыми» атаками, особенно на задачах классификации изображений, при этом оставаясь применимым в условиях ограниченного доступа к модели.

3. Методы защиты

Метод защиты на основе JPEG-сжатия основан на идее удаления из изображения высокочастотных компонентов, не воспринимаемых человеком [Das et al., 2017]. В JPEG-сжатии изображение делится на блоки (обычно 8×8) и преобразуется с помощью дискретного косинусного преобразования с последующим квантованием коэффициентов, что ослабляет высокочастотные составляющие. Так как состязательный шум чаще всего проявляется именно в виде незначимых высокочастотных колебаний,

JPEG-сжатие действует как фильтр, ослабляющий такие возмущения. Экспериментально подтверждена его эффективность: увеличение степени сжатия приводит к «восстановлению» точности на атакованных изображениях [Aydemir et al., 2024]. Однако сильное сжатие ухудшает качество изображений (блоковые артефакты, искажение цветности), что может негативно сказываться на распознавании.

Наложение шума на изображение также является эффективным методом защиты нейросетевых моделей от состязательных атак [Hossain et al., 2022]. Идея метода заключается в следующем. Так как состязательные искажения представляют собой малые возмущения, незаметные для человеческого зрения, наложение более интенсивного случайного шума способно «подавить» их влияние, снижая вероятность некорректного распознавания. Метод зашумления может быть реализован как на стадии предобработки входных изображений, так и в процессе обучения модели в рамках стратегии расширения обучающей выборки. Однако чрезмерное усиление шумового компонента способно привести к искажению семантически значимой информации. Таким образом, данный метод требует выбора оптимального уровня шума.

Нейросетевая очистка входных данных является методом защиты от состязательных атак, в котором используются специальные нейронные архитектур, способные устранять нежелательные искажения до подачи изображения в основную модель классификации. В качестве таких нейросетей обычно применяются автоэнкодеры, которые обучены восстанавливать исходное изображение на основе его искаженной версии.

Наиболее популярными являются следующие типы автоэнкодеров: автоэнкодер с шумоподавлением (Denoising Autoencoder, DAE); глубокий автоэнкодер на базе сверточных и транспонированных сверточных слоев (Convolution Autoencoder, CAE); вариационный автоэнкодер (Variational Autoencoder, VAE).

Автоэнкодеры DAE обучаются на парах (x_{noisy}, x) , где x_{noisy} – искаженный вариант изображения, а x – его «чистый» аналог. Задача автоэнкодера DAE сводится к восстановлению исходного изображения по его искаженной версии [Chen et al., 2024]. Это приводит к обучению DAE с целью минимизации потерь восстановления.

Автоэнкодеры CAE благодаря своей глубокой структуре и обучению на больших выборках могут выявлять локальные и глобальные закономерности, что позволяет повысить точность классификации после атаки на 10–20 % по сравнению с необработанными входами [Ashraf et al., 2024].

Автоэнкодеры VAE моделируют вероятностное распределение в латентном пространстве [Xie et al., 2021]. Вместо отображения входа в фиксированный вектор признаков, энкодер формирует два параметра: среднее $\mu(x)$ и дисперсию $\sigma(x)$, задающие нормальное распределение. Из этого распределения далее выбирается латентный вектор. Благодаря

регуляризации латентного пространства, VAE обеспечивает устойчивость к мелким искажениям входных данных. Он не только восстанавливает семантику изображения, но и ограничивает возможные вариации в латентной области, что делает затруднительным прохождение состязательных примеров через декодер без потерь. Таким образом, VAE может рассматриваться как робастный автоэнкодер, обладающий потенциалом к устранению состязательных возмущений.

4. Реализация

Предложенный подход был реализован в среде PyCharm. Были установлены следующие библиотеки: `import torch; import torch.nn as nn; import torch.optim as optim; import matplotlib.pyplot as plt; import numpy as np.`

В качестве набора данных изображений использовался датасет STL-10 (Stanford TensorLab). Он применяется для задач классификации цветных изображений и содержит 30000 снимков размером 96 x 96 пикселей, распределенных по 10 классам. Изображения 96 x 96 находят практическое применение в системах распознавания объектов на краевых устройствах, таких как камеры видеонаблюдения или смартфоны, где важны скорость обработки и низкое энергопотребление. Этот размер часто используется в промышленной автоматизации для классификации дефектов на конвейере – например, различия между деталями, браком или упаковкой – без необходимости в высоком разрешении. Также 96×96 активно применяется в медиасервисах для предварительной сортировки изображений, например, в фильтрации контента или тегировании фото в облачных хранилищах, где обрабатываются тысячи изображений в реальном времени.

Для распознавания изображений использовалась предобученная модель ResNet-18, адаптированная для задачи классификации на наборе данных STL-10. В первом сверточном слое conv1 проведена замена оригинальных параметров `kernel_size=7`, `stride=2`, `padding=3` на `kernel_size=3`, `stride=1`, `padding=1`. Это позволяет лучше обрабатывать изображения меньшего размера (96×96 пикселей), характерные для STL-10, сохраняя при этом пространственную информацию и уменьшая агрессивное уменьшение размерности на первом шаге. Последний полносвязный слой fc заменен с 1000 выходов на слой с 10 выходами, соответствующими количеству классов в STL-10. После этих модификаций модель дообучается на обучающей выборке STL-10, используя оптимизатор Adam и функцию потерь – кросс-энтропию. Обучающая выборка состояла из 5000 изображений, а тестовая – из 8000.

Автоэнкодеры имеют схожие архитектуры: начинаются с входного слоя, который принимает изображения размером 96x96x3. В энкодере используется архитектура, включающая два сверточных слоя. Первый сверточный слой имеет 32 фильтра с размером ядра 3x3 и функцией акти-

вации ReLU. За ним следует слой MaxPooling с размером пулинга 2x2. Второй сверточный слой содержит 64 фильтра, также с размером ядра 3x3 и активацией ReLU, после чего применяется еще один слой MaxPooling с размером пулинга 2x2.

Данный подход, объединяющий методы защиты на основе автоэнкодеров, наложения шума и JPEG-сжатия, не рассматривался в контексте систем обработки изображений в реальном времени (real-time). Автоэнкодеры, и JPEG-сжатие – являются вычислительно затратными операциями, особенно при их последовательном применении. Это приводит к увеличению общей задержки на кадр, что делает сложно достижимым соблюдение временных ограничений, характерных для real-time систем, таких как видеонаблюдение, автономные транспортные средства или интерактивные приложения.

5. Проведение экспериментов

Общий план проведения экспериментов заключался в следующем: применение каждого метода защиты в отдельности; комбинация двух методов; комбинация лучших значений всех трех методов. Определялась точность распознавания изображений из датасета STL-10 после воздействия атак FGSM, IFGSM и ZOO по следующей формуле:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \quad (5.1)$$

где TP – число правильных решений о наличии объекта; FP – число ошибок второго рода; FN – число ошибок первого рода; TN – число правильных решений об отсутствии объекта.

В табл. 1 показаны значения *Accuracy* при применения каждого метода защиты в отдельности. Значение ϵ для каждой атаки было равно 0,04 или 0,06. Значение *Accuracy* без атаки было равно 0,75.

Таблица 1

Метод защиты	FGSM		IFGSM		ZOO	
	0,04	0,06	0,04	0,06	0,04	0,06
DAE	0,26	0,19	0,39	0,34	0,66	0,62
CAE	0,35	0,29	0,40	0,36	0,49	0,51
VAE	0,21	0,12	0,23	0,13	0,64	0,64
JPEG ($q = 5$)	0,37	0,32	0,41	0,38	0,52	0,51
JPEG ($q = 10$)	0,36	0,25	0,41	0,36	0,67	0,62
JPEG ($q = 15$)	0,30	0,20	0,37	0,29	0,64	0,69
Гаусс ($\sigma = 0,2$)	0,18	0,16	0,19	0,18	0,16	0,16
Гаусс ($\sigma = 0,3$)	0,13	0,12	0,12	0,12	0,13	0,13
Гаусс ($\sigma = 0,4$)	0,11	0,11	0,11	0,11	0,09	0,09

Из рис. 1 видно, что для атак FGSM и IFGSM наилучшую точность распознавания обеспечивал автоэнкодер CAE. Для атаки ZOO при $\epsilon = 0,04$ лучшим был DAE, а при $\epsilon = 0,06$ – VAE. Наилучшим параметром сжатия JPEG для атак FGSM и IFGSM был $q = 5$, а для ZOO – $q = 10$ при $\epsilon = 0,04$ и $q = 15$ при $\epsilon = 0,06$. Наилучшим параметром шума Гаусса для всех атак – $\sigma = 0,2$, хотя влияние шума на защиту от атак было наименее сильным.

При применении попарной комбинации методов оптимальные параметры методов для различных атак могут становиться другими. В табл. 2 показаны значения *Accuracy* с учетом этих оптимальных параметров.

Таблица 2

Комбинация методов защиты	FGSM		IFGSM		ZOO	
	0,04	0,06	0,04	0,06	0,04	0,06
DAE + JPEG ($q = 5$)	0,42	0,37	-	-	-	-
CAE + JPEG ($q = 10$)	0,36	0,32	0,48	0,44	0,57	0,56
CAE + JPEG ($q = 15$)	0,36	0,32	0,46	0,42	0,62	0,64
DAE + Гаусс ($\sigma = 0,2$)	0,30	0,23	0,38	0,35	0,47	0,49
JPEG ($q = 10$) + Гаусс ($\sigma = 0,2$)	0,37	0,36	0,38	0,35	0,49	0,31
JPEG ($q = 15$) + Гаусс ($\sigma = 0,2$)	0,36	0,30	0,35	0,30	0,48	0,46

Из сравнения табл. 1 и 2 видно, что значения *Accuracy* при попарной комбинации (AE + JPEG) выше, чем при одиночном использовании этих методов при тех же параметрах. Если в попарной комбинации используется метод зашумления, то значения *Accuracy* будут выше, чем при одиночном использовании метода зашумления, но ниже, чем при одиночном использовании метода нейронной очистки или JPEG-сжатия.

Последним этапом проведения экспериментов являлось полное комбинирование всех трех методов защиты (табл. 3).

Таблица 3

Комбинация методов защиты	FGSM		IFGSM		ZOO	
	0,04	0,06	0,04	0,06	0,04	0,06
DAE+JPEG($q=5$)+Гаусс($\sigma=0,2$)	0,36	0,30	0,39	0,35	0,39	0,45
DAE+JPEG ($q=10$)+Гаусс($\sigma=0,2$)	0,31	0,26	0,35	0,33	0,39	0,41
CAE+JPEG($q=5$)+Гаусс ($\sigma=0,2$)	0,23	0,21	0,27	0,25	0,32	0,29
CAE+JPEG ($q=10$)+Гаусс($\sigma=0,2$)	0,30	0,23	0,27	0,26	0,29	0,27

Как видно из табл. 3, для всех атак при различных значениях ϵ наибольшее значение *Accuracy* достигается в случае комбинирования методов DAE, JPEG-сжатия с $q = 5$ и добавления шума Гаусса с $\sigma = 0,2$. При этом, сравнивая табл. 3 и 2, можно заметить, что при полном комбинировании методов защиты значения *Accuracy* меньше, чем при парном применении автоэнкодеров и JPEG-сжатия, меньше, чем при парном приме-

нении автоэнкодеров и зашумления и примерно равны при парном применении JPEG-сжатия и зашумления. Это объясняется особенностями применения метода зашумления. По-видимому, этот метод, помимо противодействия состязательным атакам, все-таки оказывает негативное воздействие и на другие методы защиты. Таким образом, проведенные исследования показывают, что наибольшей эффективностью противодействия градиентным состязательным атакам обладает комбинация методов нейронной очистки и JPEG-сжатия.

Заключение

В работе исследован подход к защите систем распознавания изображений от градиентных состязательных атак, основанный на комбинировании нескольких методов защиты, в качестве которых рассматривались методы нейронной очистки с помощью автоэнкодеров, JPEG-сжатия и наложения шумов Гаусса. При этом были найдены оптимальные значения параметров методов защиты, при которых достигается наивысшая точность распознавания изображений.

Исследование показало, что наиболее эффективным является способ попарного применения нейронной очистки и JPEG-сжатия. Способ полного комбинирования методов защиты уступает по эффективности из-за негативного эффекта метода зашумления, который влияет не только на состязательные атаки, но и на другие методы защиты.

Предложенный в работе подход обладает потенциалом для дальнейшего развития. Планируется рассмотреть другие комбинации защитных механизмов, применять различные архитектуры нейронных сетей для обработки изображений, протестировать подход на дополнительных датасетах и более сложных моделях классификации, а также провести сравнение предложенного метода с гибридными методами.

Список литературы

- [Баев и др., 2024] Баев А.В., Самонов А.В., Сафонов В.М., Краснов С.В., Малышев С.Р. Методы защиты моделей нейронных сетей от состязательных атак уклонения и отравления // Автоматизация процессов управления. – 2024. – № 4 (78). – С. 39-48. – doi: 10.35752/1991-2927_2024_4_78_39.
- [Котенко и др., 2025] Котенко И.В., Саенко И.Б., Лаута О.С., Васильев Н.А., Садовников В.Е. Метод противодействия состязательным атакам на системы классификации изображений // Вопросы кибербезопасности. – 2025. – № 2(66). – С. 114-123. – doi: 10.21681/2311-3456-2025-2-114-123.
- [Легашев и др., 2024] Легашев Л.В., Жигалов А.Ю. Исследование состязательных атак на регрессионные модели машинного обучения в беспроводных сетях 5G // Вопросы кибербезопасности. – 2024. – № 3(61). – С. 61-67. – doi: 10.21681/2311-3456-2024-3-61-67.

- [Понкин и др., 2024] Понкин И.В., Куприяновский В.П., Морева С.Л., Лаптева А.И. Компьютерное зрение: концепт, функционально-целевое назначение, структура, регуляторика // International Journal of Open Information Technologies. – 2024. – Т. 12, № 5. – С. 57-66.
- [Фомичева и др., 2023] Фомичева С.Г., Беззатеев С.В. Механизмы защиты моделей машинного обучения от состязательных атак // T-Comm: Телекоммуникации и транспорт. – 2023. – Т. 17, № 10. – С. 28-42. – doi: 10.36724/2072-8735-2023-17-10-28-42.
- [Ali et al., 2022] Ali Y., Wani M.A. HAT: Hybrid Adversarial Training to Make Robust Deep Learning Classifiers // In: Proc. 2022 9th International Conference on Computing for Sustainable Global Development (INDIACom). – 2022. – P. 433-436. – doi: 10.23919/INDIACom54597.2022.9763284.
- [Ashraf et al., 2024] Ashraf S.N., Siddiqi R., Farooq H. Auto encoder-based defense mechanism against popular adversarial attacks in deep learning // PLoS ONE. – 2024. – Vol. 19(10). – P. e0307363. doi: 10.1371/journal.pone.0307363.
- [Aydemir et al., 2024] Aydemir A.E., Temizel A., Temizel T.T. The effects of JPEG and JPEG2000 compression on attacks using adversarial examples. 2024. arXiv:1803.10418.
- [Chen et al., 2017] Chen P.Y., Zhang H., Sharma Y., Yi J., Hsieh Ch.-J. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models // In: Proc. 10th ACM workshop on artificial intelligence and security (AISec '17). – 2017. – P. 15-26. – doi: 10.1145/3128572.3140448.
- [Chen et al., 2024] Chen Z., Chen Q., Zhou H., Zhang H. DAE-Net: Deforming auto-encoder for fine-grained shape co-segmentation // In: Proc. ACM SIGGRAPH 2024 Conference Papers (SIGGRAPH '24), Denver, CO USA, 2024. Article No. 82. – P. 1-11. – doi: 10.1145/3641519.3657528.
- [Das et al., 2017] Das N., Shanbhogue Мюб., Chen Sh.-T., Hohman F., Chen L., Kounavis M.E., Chau D.H. Keeping the bad guys out: Protecting and vaccinating deep learning with JPEG compression. – 2017. – arXiv:1705.02900.
- [Gardella et al., 2022] Gardella M., Nikoukhah T., Li Y., Bammey Q. The Impact of JPEG Compression on Prior Image Noise // In: Proc. ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022. – P. 2689-2693. – doi: 10.1109/ICASSP43922.2022.9746060.
- [Goodfellow et al., 2015] Goodfellow I.J., Shlens J., Szegedy C. Explaining and harnessing adversarial examples. – 2015. – arXiv:1412.6572.
- [Hossain et al., 2022] Hossain M.T., Badsha S., La H., Islam S., Khalil I. Exploiting Gaussian noise variance for dynamic differential poisoning in federated learning // IEEE Transactions on Artificial Intelligence. – 2022. – Vol. 1, No. 01. – P. 1-17. – doi: 10.1109/TAI.2025.3540030.
- [Jain, 2024] Jain S. Adversarial attack on Yolov5 for traffic and road sign detection // In: Proc. 2024 4th International Conference on Applied Artificial Intelligence (ICAPAI). – 2024. – P. 1-5. – doi: 10.1109/ICAPAI61893.2024.10541282.
- [Jiakai, 2018] Jiakai W. Adversarial examples in the physical world. Artificial intelligence safety and security // In: Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21). – 2018. – P. 4925-4926. – doi: 10.24963/ijcai.2021/694.

- [**Khamaiseh et al., 2022**] Khamaiseh S.Y., Bagagem D., Al-Alaj A., Mancino M., Alomari H.W. Adversarial deep learning: A survey on adversarial attacks and defense mechanisms on image classification // IEEE Access. – 2022. – Vol. 10. – P. 102266-102291. – doi: 10.1109/ACCESS.2022.3208131.
- [**Kotenko et al., 2024**] Kotenko I., Saenko I., Laut O., Vasiliev N., Sadovnikov V. An approach to countering adversarial attacks on image recognition based on JPEG-compression and Neural-Cleanse // In: Proc. 2024 IEEE Ural-Siberian Conference on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT). – 2024. – P. 76-79. – doi: 0.1109/USBEREIT61901.2024.10584049.
- [**Kotlyarov et al., 2023**] Kotlyarov D.V., Dyudyun G.D., Rzhhevskaya N.V., Lapina M.A., Babenko M.G. Investigation of adversarial attacks on pattern recognition neural networks // Proceedings of the Institute for System Programming of the RAS. – 2023. – Vol. 35, No. 2. – P. 35-48. – doi: 10.15514/ISPRAS-2023-35(2)-3.
- [**Li et al., 2023**] Li B., Wu S., Yang Y., Zhang G. Analysis and research of neural network adversarial samples for power grid security // In: Proc. 2023 8th International Conference on Data Science in Cyberspace (DSC). – 2023. – P. 520-525, – doi: 10.1109/DSC59305.2023.00081.
- [**Naveen et al., 2023**] Naveen I.G., Inchara, Meghana H., Preethi N., Neha B. A combined approach for efficient compression and restoration of multispectral satellite images // In: Proc. 2023 International Conference on Applied Intelligence and Sustainable Computing (ICAISC). – 2023. – P. 1-5. – doi: 10.1109/ICAISC58445.2023.10200770.
- [**Tian et al., 2024**] Tian P., Poreddy S., Danda C., Gowrineni C., Wu Y., Liao W. Evaluating impact of image transformations on adversarial examples // IEEE Access. – 2024. – Vol. 12. – P. 186217-186228. – doi: 10.1109/ACCESS.2024.3487479.
- [**Xie et al., 2021**] Xie Z., Liu C., Zhang Y., Lu H., Wang D., Ding, Y. Adversarial and contrastive variational autoencoder for sequential recommendation // In: Proc. Web Conference 2021 (WWW'21). – 2021. – P. 449-459. – doi: 10.1145/3442381.3449873.