МЕТОД ПРОТИВОДЕЙСТВИЯ СОСТЯЗАТЕЛЬНЫМ АТАКАМ НА СИСТЕМЫ КЛАССИФИКАЦИИ ИЗОБРАЖЕНИЙ

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

DOI: 10.21681/2311-3456-2025-2-114-123

Цель исследования: *разработка и оценка метода противодействия состязательным атакам FGSM, ZOO, OPA на системы классификации изображений, основанного на интеграции процедур зашумления, нейронной очистки и JPEG-сжатия данных.*

Методы исследования: *системный анализ, машинное обучение, зашумление изображения, нейронная очистка, JPEG-сжатие данных, вычислительный эксперимент.*

Полученные результаты: *проведен анализ работ по тематике атак на системы классификации изображений (СКИ), основанные на применении методов машинного обучения, и методов защиты от них. По результатам проведенного анализа выявлено, что к числу наиболее часто встречающихся атак на СКИ относятся состязательные атаки, а именно: быстрый метод градиентного знака (FGSM), оптимизация нулевого порядка (ZOO) и атака одним пикселем (OPA). Тематика противодействия этим атакам в настоящее время вызывает повышенный интерес. Раскрыта сущность воздействия этих атак на СКИ, а также выявлены их влияние на точность распознавания изображений. Предложен метод противодействия состязательным атакам, основанный на зашумлении изображения шумами Гаусса и Пуассона, а также применении JPEG-сжатия и технологии нейронной очистки. Проведены эксперименты, показывающие высокую эффективность предложенного метода. Эксперименты были направлены на оценку точности распознавания изображений, содержащихся в двух различных наборах данных – наборе изображений деталей персонального компьютера и наборе рукописных цифровых изображений. Оценивались результаты распознавания изображений до и после воздействия на СКИ состязательных атак, а также после применения к этим наборам предложенного метода.*

Научная новизна: *анализ работ по тематике защиты от состязательных атак показал, что в настоящее время наиболее характерными атаками на СКИ являются атаки FGSM, ZOO и OPA. Предложенный метод противодействия состязательным атакам на СКИ отличается от других известных методов защиты тем, что он интегрирует возможности противодействия атакам, содержащиеся в трех различных подходах (нейронной очистке, зашумлении и JPEG-сжатии) и выявляет оптимальные параметры этих подходов. Высокая эффективность предложенного метода была подтверждена в экспериментах, проведенных на двух разных наборах данных.*

Вклад: *Котенко И. В. и Саенко И. Б. – общая концепция состязательных атак на СКИ и методов защиты от них на основе известных работ; Котенко И. В. и Лаута О. С. – описание методов воздействия состязательных атак; Васильев Н. А. и Садовников В. Е. – реализация предложенного подхода; Котенко И. В. и Саенко И. Б. – теоретическое обоснование предложенного подхода.*

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

Введение

Искусственный интеллект и машинное обучение (МО) все чаще используются для создания систем поддержки принятия решений, которые могут обрабатывать и анализировать информацию так же, как это делают люди. Суть работы систем МО заключается в их способности самостоятельно обучаться на основе больших объемов экспериментальных данных и применять полученные знания для принятия решений в новых ситуациях. Эффективность систем МО обеспечивается их способностью быстро обнаруживать скрытые закономерности в данных, описывающих различные процессы или явления.

Системы МО достигли значительных успехов в различных областях применения, например, в распознавании речи [1], машинном переводе [2], автономном управлении транспортными средствами [3],

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

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

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

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

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

медицине [4]. Особую значимость МО приобретает в системах классификации изображений (СКИ), так как они являются основой для многих приложений в различных сферах науки и техники [5–7]. Однако, несмотря на значительное повышение точности МО, недавние исследования показали, что системы МО, в том числе СКИ, могут быть очень уязвимы для особого типа атак, называемых состязательными. Они представляют собой специально разработанные входные данные, которые могут ввести нейронную сеть в заблуждение и привести к неверным результатам. Примером в задаче классификации изображений является случай, когда естественное изображение намеренно искажается визуально незаметными изменениями, приводящими к резкому снижению точности классификации [8]. Эти атаки могут быть использованы, например, для нарушения работы таких разновидностей СКИ, как системы распознавания лиц или системы распознавания дорожных знаков [9].

Одними из наиболее распространенных и опасных состязательных атак являются следующие: быстрый метод градиентного знака (Fast Gradient Sign Method, FGSM), оптимизация нулевого порядка (Zeroth Order Optimization, ZOO) и атака одним пикселем (One Pixel Attack, OPA) [10]. Эти атаки могут серьезно нарушить работу нейронных сетей и привести к неверным результатам, поскольку они манипулируют входными данными с использованием высокочастотных компонентов, которые человек не может обнаружить.

В настоящей статье мы рассмотрим меры противодействия этим атакам, основанные на добавлении гауссовского и пуассоновского шума, а также технологий нейронной очистки (Neural Cleanse) и JPEG-сжатия. Добавление шума к изображению направлено на то, чтобы исказить действие вражеской атаки и направить ее в неверном направлении. Технология нейронной очистки является хорошо известным методом обнаружения и смягчения последствий агрессивных атак со стороны нейронных сетей. Для обнаружения и удаления изображений, содержащих агрессивные атаки, используется алгоритм контролируемого обучения. Сжатие JPEG – это известный метод удаления высокочастотных компонентов из защищенных изображений, который также может предотвратить успешное проведение состязательных атак.

В своих предыдущих исследованиях мы частично исследовали возможность использования технологий нейронной очистки и сжатия JPEG [11] и убедились в их эффективности. В настоящей работе мы предлагаем дополнить эти меры противодействия подходом, основанным на добавлении к изображениям двух типов шума (Гаусса и Пуассона), и найти оптимальные параметры этих шумов, обеспечивающие максимальную точность классификации изображений после воздействия состязательных атак. Это определяет цель, новизну и вклад нашей работы.

Анализ работ по тематике атак на СКИ и методов защиты от них

Было проанализировано множество работ по созданию состязательных атак и разработке соответствующих методов защиты от них СКИ. В целом состязательные атаки могут быть классифицированы как атаки «белого ящика» и «черного ящика» на основе информации о модели цели, доступной злоумышленнику. При атаке «белым ящиком» злоумышленник имеет полный доступ к сетевой архитектуре и параметрам, в то время как при атаках «черным ящиком» разрешен только внешний доступ к сети (например, ввод и вывод данных).

В работе [12] авторы предлагают метод противодействия состязательным атакам, основанный на использовании комплекса различных преобразований изображений. Основная идея метода заключается в том, что преобразования изображений, такие как добавление шума, изменение яркости и контрастности, поворот и масштабирование, могут изменить входные данные (состязательные выборки) таким образом, что они больше не будут успешными. Более того, если вы используете только одно преобразование, то злоумышленник может адаптироваться к нему и создать новые образцы противодействия, которые будут успешными. В ходе экспериментов авторы продемонстрировали, что этот метод эффективен против многих типов состязательных атак, включая атаку FGSM.

В работе [13] представлен обзор состязательных атак и методов защиты в области глубокого обучения. Подчеркивается, что состязательные атаки приводят к изменениям во входных данных, которые обычно невидимы для человеческого восприятия, но могут существенно повлиять на эффективность моделей глубокого обучения для классификации объектов. В статье рассматриваются различные типы состязательных атак, включая FGSM, ZOO. Кроме того, рассматриваются преимущества и недостатки различных подходов к защите от состязательных атак, таких как увеличение обучающих данных, стеганография, перегонка моделей, защитные нейронные сети, состязательные модели и другие.

В работе [14] рассматривается возможность использования JPEG-сжатия для противодействия состязательным атакам на глубокие нейронные сети. Идея метода заключается в том, что сжатие JPEG, которое широко используется для сжатия изображений, может разрушать высокочастотные компоненты изображения, которые часто используются в атаках злоумышленников. В то же время сжатие в формате

JPEG не влияет на качество изображения, воспринимаемого человеком, и может быть легко реализовано. В ходе экспериментов авторы продемонстрировали, что их метод эффективен против различных типов состязательных атак, включая FGSM и OPA. Было показано, что этот метод может быть успешно применен для защиты различных типов нейронных сетей, образующих основу СКИ.

В работе [15] авторы используют метод обнаружения и защиты от враждебных атак на основе нейронных сетей, названный нейронной очисткой. Идея этого метода заключается в том, что состязательные атаки оставляют следы в нейронной сети, которые могут быть обнаружены и удалены. Чтобы обнаружить эти следы, предлагается использовать градиентное обратное распространение, чтобы определить, какие нейроны в сети активируются наиболее сильно при обработке состязательных примеров. Затем, используя этот набор нейронов, создается специальный детектор, который может обнаруживать состязательные атаки. В ходе экспериментов авторы продемонстрировали, что метод нейронной очистки эффективен против атак типа FGSM и некоторых других.

В работе [16] рассматриваются атаки, которые могут быть использованы для обмана моделей глубокого обучения, таких как нейронные сети. Эти атаки предполагают внесение незначительных изменений во входных данных, которые незаметны для человека, но могут привести к ошибочным прогнозам модели. Кроме того, рассматриваются различные методы защиты от подобных атак, которые направлены на совершенствование архитектуры моделей МО, а также разработку специальных алгоритмов для обнаружения и нейтрализации состязательных примеров.

Таким образом, анализ работ по тематике атак на СКИ и методов защиты от них показывает, что известен ряд подходов к противодействию состязательным атакам, в которых используются нейронная очистка и сжатие JPEG. Однако подходы, основанные на добавлении шума, изучены слабо. Наше исследование направлено на устранение этого недостатка.

Характеристика атак FGSM, ZOO и OPA

Выбор атак FGSM, ZOO и OPA обусловлен тем, что, во-первых, они относятся к разным классам «белого» и «черного ящиков» (атака FGSM – это атака «белого ящика», а атаки ZOO и OPA – это атаки «черного ящика»), а с другой стороны, – что они, в основном, ориентированы на СКИ.

FGSM – это метод атаки на нейронные сети, который используется для введения в заблуждение обученных моделей. Суть данного метода заключается в незначительном изменении изображения, чтобы обученная модель неправильно идентифицировала его как другой класс. Для этого используется метод градиентного спуска, который позволяет обнаружить наиболее чувствительные пиксели на изображении [17].

При использовании метода FGSM исходное изображение рассматривается как точка на пути от исходного к измененному изображению, что обеспечивает максимальное изменение значения целевой функции потерь (Loss Function). Затем вычисляется градиент потерь относительно каждого пикселя изображения. После этого все пиксели с наименьшим модулем градиента обнуляются, а остальные увеличиваются или уменьшаются на величину, соответствующую знаку градиента [18]. Метод FGSM позволяет создавать поддельные изображения, которые практически неотличимы от оригиналов, но содержат измененную информацию, которая может ввести в заблуждение модель машинного обучения (рис. 1).

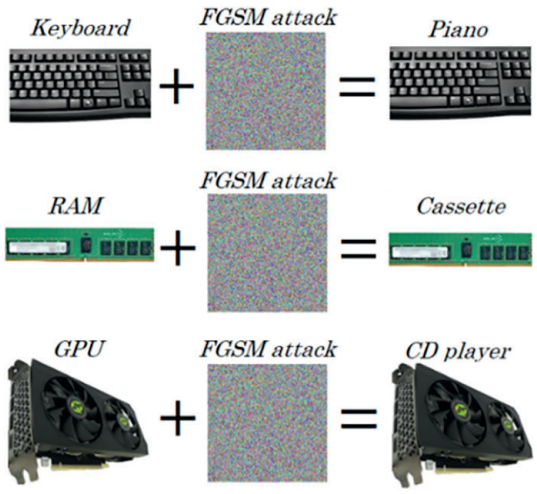

Рис. 1. Влияние атаки FGSM на распознавание изображений

Атака ZOO – это метод оптимизации, применяемый в МО, который не требует информации о градиенте целевой функции, а использует только значения целевой функции для ее оптимизации [19]. Он работает путем добавления небольших возмущений к исходному входному изображению и наблюдения за тем, как изменяется результат модели. Основываясь на этих наблюдениях, метод ZOO оптимизирует возмущения, чтобы максимизировать вероятность неправильного предсказания модели. При таком подходе злоумышленник не имеет прямого доступа к внутренним параметрам модели и ее структурной информации. Вместо этого он взаимодействует с внешним API модели, отправляя входные данные и получая выходные, но без какого-либо доступа к внутренним параметрам (рис. 2).

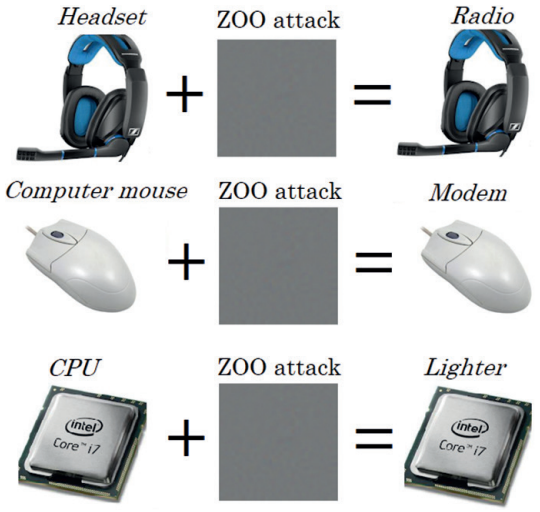

Рис. 2. Влияние атаки ZOO на распознавание изображений

OPA-атака – это атака на компьютерное зрение и обработку изображений, при которой изменяется всего один пиксель изображения таким образом, чтобы обмануть СКИ и изменить ее выходные данные (рис. 3). Эта атака заключается в поиске единственного пикселя на изображении, изменение которого приведет к максимальному изменению выходных данных СКИ, что достигается с помощью методов оптимизации, направленных на максимизацию функции потерь. Изменение в одном пикселе, с одной стороны, может быть незаметным для человеческого глаза, но с другой – существенно изменить выходные данные СКИ [20].

Рис. 3. Влияние атаки ZOO на распознавание изображений

Методы противодействия состязательным атакам

Методы противодействия враждебным атакам, рассмотренные в этой статье, включают добавление гауссовского шума и пуассоновского шума, а также использование технологий нейронной очистки и JPEG-сжатия.

Гауссовский шум – это статистический шум, который имеет простую функцию плотности вероятности, называемую функцией Гаусса или нормальным распределением. Гауссовский шум определяется следующим выражением:

$$P_G(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2\sigma^2}},$$

где μ – математическое ожидание, а σ – стандартное отклонение.

Гауссовский шум характеризуется двумя параметрами: математическим ожиданием (μ) и стандартным отклонением (σ). Математическое ожидание в гауссовском шуме определяет основную тенденцию распределения и представляет собой среднее значение на всех возможных значениях шума. Если среднее значение равно нулю, то это означает, что положительные и отрицательные значения шума в среднем уравновешиваются.

Гауссовский шум может использоваться в контексте защиты СКИ для повышения их устойчивости к атакам, основанным на повреждении входных данных. Это достигается путем добавления гауссовского шума к входным изображениям во время обучения системы. Благодаря этому система учится распознавать изображения, содержащие шум, и становится менее чувствительной к искажениям во входных данных.

Пуассоновский шум – это стохастический шум, который возникает из-за неконтролируемого распределения случайных событий, таких как попадание фотографий на матрицу изображения в цифровых камерах. Он характеризуется тем, что его амплитуда зависит от среднего количества событий, происходящих в данном пространстве. Математически пуассоновский шум описывается распределением Пуассона, которое задается следующей функцией вероятности:

$$p(k) = \frac{\lambda^k}{k!} e^{-\lambda},$$

где k – количество событий, произошедших за данный промежуток времени или пространства, λ – средняя скорость (интенсивность) событий (случайная величина).

Метод нейронной очистки состоит из двух основных этапов: обучения и перебора. На первом этапе модель обучается на исходном наборе данных. Затем, на втором этапе, модель оценивает входные данные, чтобы определить, соответствуют ли они экземпляру вредоносного программного обеспечения. Основная концепция метода нейронной очистки заключается в том, что существует связь между входными и выходными данными модели. Для достижения этой цели используется показатель, называемый «доверительные веса», который измеряет уровень достоверности прогнозов модели. После определения доверительных весов используется алгоритм оптимизации для поиска точек данных с наибольшими весами. Этот процесс позволяет обнаруживать состязательные примеры и удалять их из обучающего набора данных.

JPEG-сжатие является одним из наиболее распространенных методов сжатия изображений, который используется для уменьшения размера графических файлов. JPEG-сжатие является сжатием с потерями, что означает потерю части информации об изображении. Этот метод сжатия является неотъемлемой частью процесса уменьшения объема передаваемых данных и повышения эффективности СКИ. Для защиты СКИ важно соблюдать баланс между сжатием изображений для экономии ресурсов и поддержанием достаточного качества изображения для точного распознавания объектов.

Реализация подхода

Предложенный подход был реализован на следующих двух наборах данных:

- набор изображений деталей персонального компьютера (ПК) PC Parts Images (PCPI), взятый из базы данных Kaggle[6]. Этот набор данных состоит из 3279 изображений компонентов ПК. Набор данных содержит 14 классов. Каждое изображение имеет разрешение 256x256 пикселей и выполнено в формате JPG;
- набор данных MNIST-JPG[7]. Это выборка из хорошо известного набора данных MNIST. Набор данных MNIST содержит 60 000 обучающих изображений и 10 000 тестовых, каждое из которых представляет собой изображение в оттенках серого размером 28x28 пикселей. Набор данных MNIST-JPG состоит из 1000 рукописных изображений чисел, преобразованных в формат JPG. Каждое число в наборе данных MNIST-JPG имеет 100 различных вариантов написания.

Был определен набор классификаторов, которые использовались для распознавания изображений, который включает в себя следующие классификаторы: k-ближайших соседей (KNN), случайный лес (RF), наивный Байесовский алгоритм (NB), деревья решений (DT). Выбор этих классификаторов обусловлен тем, что они получили достаточно широкое распространение в современных СКИ. Параметры всех классификаторов были подобраны таким образом, чтобы их функция потерь при обучении уменьшалась, а точность повышалась.

Предложенный подход был реализован в среде PyCharm. Были установлены следующие библиотеки: import matplotlib.pyplot as plt, import numpy as np, import os, from art.utils import load_dataset, from art.attacks.evasion import HopSkipJump, from art.estimators.classification import SklearnClassifier.

Построение графика проводилось с помощью модуля Matplotlib.

Общая процедура оценки эффективности подхода была одинаковой для каждого набора данных. Она включала в себя следующие шаги.

Шаг 1: обучение выбранных классификаторов на основе исходного набора данных. Используется 80 % записей. Применяется 10-кратная перекрестная проверка.

Шаг 2: тестирование распознавания изображений на исходном наборе данных. Здесь используется показатель F-scope, который вычисляется как $F = 2TP / (2TP + FP + FN)$, где TP – истинно положительный результат, FP – ложноположительный результат, FN – ложноотрицательный результат. Значение F оценивает эффективность распознавания изображений перед атакой и ниже обозначается как F_0.

Шаг 3: создание трех копий исходного набора данных и дальнейшее проведение атак FGSM[8], ZOO[9] и OPA[10]. Для этой цели был использован код на языке Python.

Шаг 4: оценка распознавание изображений после атаки, используя метрику F-scope, которая оценивает эффективность распознавания изображений после атаки. Это значение ниже обозначается как F_1.

Шаг 5: применение шума к каждому изображению исходного набора данных с помощью кода, написанного на языке Python.

Шаг 6: повторение шага 2 для наборов данных с шумом, т.е. подверженных атакам FGSM, ZOO и OPA.

Шаг 7: применение технологии нейронной очистки и JPEG-сжатия ко всем изображениям, подвергшимся атаке.

Шаг 8: оценка распознавания изображений с помощью показателя F-scope, который оценивает эффективность распознавания изображений после применения нашего подхода. Он ниже будет обозначен как F_2.

Эксперименты и обсуждение

В первой части экспериментов использовался набор данных PCPI. На нем была оценена эффективность предлагаемого подхода к противодействию атакам FGSM, ZOO, OPA и были определены оптимальные параметры гауссовского и пуассоновского шума для противодействия этим атакам. Искомыми параметрами шумов являлись стандартное отклонение гауссова шума и среднее значение пуассоновского шума.

В таблице 1 представлены показатели эффективности распознавания изображений в наборе данных PCPI до (F_0) и после (F_1) атак FGSM, ZOO и OPA.

6 https://www.kaggle.com/datasets/asaniczka/pc-parts-images-dataset-classification ?resource=download
7 https://github.com/teavanist/MNIST-JPG
8 https://github.com/1Konny/FGSM
9 https://github.com/IBM/ZOO-Attack
10 https://github.com/drexpp/Adversarial-examples-One-pixel-attack

Таблица 1.

Эффективность распознавания изображений PCPI до и после атак FGSM, ZOO и OPA

Классификатор	До атаки (F_0)	После атаки (F_1)		
		FGSM	ZOO	OPA
RNN	0,698	0,145	0,182	0,695
RF	0,878	0,311	0,321	0,880
NB	0,789	0,184	0,209	0,787
DT	0,756	0,167	0,244	0,754

Как видно из таблицы 1, для методов FGSM и ZOO эффективность распознавания после атаки значительно снижена для всех классификаторов по сравнению со случаем до атаки. Кроме того, RF-классификатор демонстрирует наивысшую эффективность распознавания до атаки, но после атаки его эффективность снижается в большей степени, чем у других классификаторов. Таким образом, RF-классификатор является наиболее уязвимым.

При этом для атаки OPA эффективность распознавания всех классификаторов остается примерно такой же, как и до атаки. Из этого можно сделать вывод, что атака OPA снижает эффективность распознавания изображений на уровне ошибок. Другими словами, OPA не оказывает существенного влияния на эффективность распознавания изображений в наборе данных PCPI. Это может быть связано с тем, что каждое изображение в наборе данных PCPI имеет разрешение 256x256 пикселей. Этот вывод подтверждается данными, представленными на рисунках 4 и 5, на которых показано влияние атак ZOO и OPA на распознавание изображений из набора данных PCPI для классификатора RF.

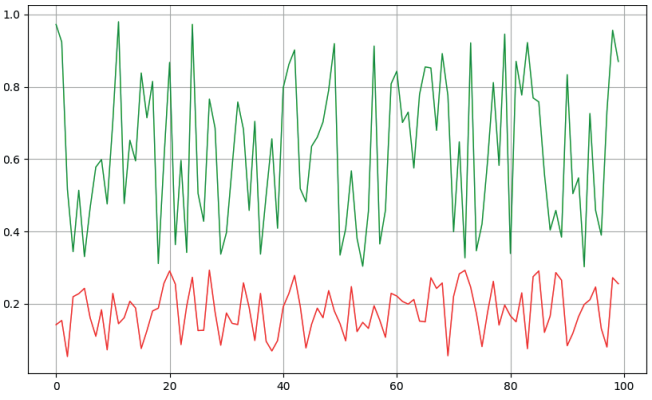

Рис. 4. Влияние атаки ZOO на распознавание изображений PCPI для RF-классификатора

Результаты классификации изображений до атаки представлены зелеными (рис. 4) или синими (рис. 5) линиями, а после атаки – красными линиями. На графиках ось X отражает номера итераций классификаторов, а ось Y представляет точность распознавания цифровых изображений, которые подаются на вход классификатора. Изображения компонентов персонального компьютера были случайным образом выбраны из одного типа в количестве 100 изображений. Эффективность распознавания оценивалась на основе количества правильно идентифицированных компонентов персонального компьютера.

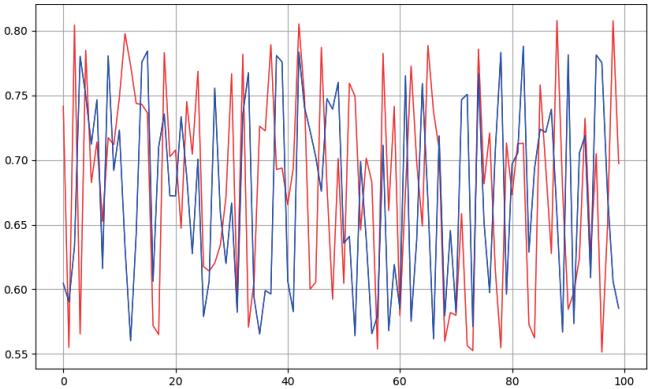

Рис. 5. Влияние атаки OPA на распознавание изображений PCPI для RF-классификатора

Легко заметить, что атака ZOO оказывает значительное влияние на распознавание изображений, содержащихся в наборе данных PCPI. Влияние атаки OPA, напротив, практически незначительно. Средние значения показателя эффективности распознавания, полученные из графиков, представленных на рисунке 4, отличаются друг от друга на уровне статистической погрешности.

В таблице 2 представлен результат эксперимента по определению оптимальных параметров гауссовского и пуассоновского шума для противодействия атакам FGSM и ZOO на распознавание изображений из набора данных PCPI с использованием классификатора RF, который показал наибольшую эффективность.

Таблица 2.

Эффективность распознавания изображений из набора данных PCPI в зависимости от параметров гауссовского и пуассоновского шума для RF-классификатора

No.	Шум Гаусса (σ)	Шум Пуассона (λ)	После атаки (F_1)	
			FGSM	ZOO
1	20	0,20	0,392	0,631
2	30	0,25	0,540	0,772
3	40	0,30	0,556	0,698
4	50	0,35	0,481	0,603
5	60	0,40	0,337	0,485
6	70	0,50	0,213	0,376

Эксперимент состоял из последовательного применения гауссовского и пуассоновского шума, а также последующего применения технологий нейронной очистки и сжатия JPEG. В ходе эксперимента было выявлено, что гауссовский шум при $\sigma < 20$ практически не влияет на эффективность распознавания, а при $\sigma > 70$ снижает ее по сравнению с эффективностью после атаки. Также было установлено, что при $\lambda < 0{,}2$ пуассоновский шум не оказывает влияния на эффективность распознавания, а при $\lambda > 0{,}5$, наоборот, снижает ее.

Эффективность распознавания изображений под воздействием FGSM–атаки достигает своего максимума при $\sigma = 40$ и $\lambda = 0{,}30$, а под воздействием ZOO – при $\sigma = 30$ и $\lambda = 0{,}25$.

Никаких экспериментов с OPA не проводилось, поскольку эта атака не оказывает существенного влияния на набор данных PCPI.

Вторая часть экспериментов проводилась на наборе данных MNIST-JPG. В таблице 3 представлены показатели эффективности распознавания изображений набора данных MNIST-JPG до и после атак FGSM, ZOO и OPA.

Таблица 3.
Эффективность распознавания изображений до и после атак FGSM, ZOO и OPA

Классификатор	До атаки (F_0)	После атаки (F_1)		
		FGSM	ZOO	OPA
RNN	0,723	0,171	0,190	0,412
RF	0,756	0,158	0,214	0,432
NB	0,794	0,208	0,202	0,387
DT	**0,910**	**0,315**	0,193	0,335

Следует отметить, что атака OPA на набор данных MNIST-JPG, в отличие от набора данных PCPI, также оказывает существенное влияние на эффективность распознавания изображений. Этот вывод подтверждается данными, представленными на рисунке 6, на котором показано влияние атаки OPA на распознавание изображений из набора данных MNOST-JPG для классификатора DT.

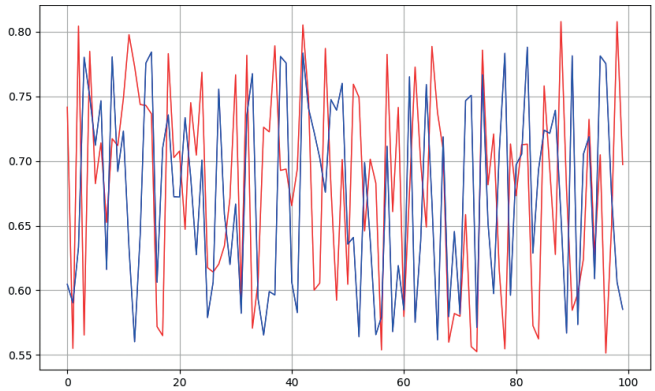

Рис. 6. Влияние OPA на распознавание изображений MNIST-JPG для классификатора DT

В таблице 4 представлены результаты эксперимента по нахождению оптимальных параметров гауссовского и пуассоновского шума для противодействия всем типам атак на распознавание изображений из набора данных MNIST-JPG при использовании классификатора DT. Как видно из таблицы 4, добавление шума позволяет восстановить эффективность распознавания изображений для всех типов атак. Максимальная эффективность распознавания изображений F2 после применения предлагаемого подхода достигается при различных параметрах шума для разных атак. Для атаки FGSM наиболее благоприятен сильный шум для противодействия атакам, для атак ZOO и OPA – наиболее благоприятен средний шум.

Таблица 4.
Эффективность распознавания изображений после применения предлагаемого подхода к классификатору DT

No.	Шум Гаусса (σ)	Шум Пуассона (λ)	После атаки (F_1)		
			FGSM	ZOO	OPA
1	20	0,20	0,173	0,214	0,434
2	30	0,25	0,199	0,284	0,446
3	**40**	**0,30**	0,284	0,402	**0,587**
4	**50**	**0,35**	0,445	**0,667**	**0,494**
5	**60**	**0,40**	**0,732**	**0,621**	0,433
6	**70**	**0,50**	**0,678**	0,325	0,326

Заключение

В статье представлен и исследуется новый подход к предотвращению состязательных атак на СКИ, который основан на применении гауссовского и пуассоновского шума в сочетании с технологиями нейронной очистки и JPEG-сжатия. Основной целью исследования была оценка влияния предложенного подхода на эффективность СКИ, использующих наборы данных PCPI и MNIST-JPG под влиянием состязательных атак, таких как FGSM, ZOO и OPA. Кроме того, целью исследования было определение оптимальных параметров гауссовского и пуассоновского шума, которые позволили бы максимально повысить устойчивость системы к этим состязательным атакам.

В ходе экспериментов было обнаружено, что каждый тип состязательной атаки наилучшим образом реагирует на определенные значения параметров гауссовского и пуассоновского шума. Кроме того,

эксперименты выявили интригующее наблюдение: атака OPA не оказала существенного влияния на эффективность распознавания изображений в наборе данных PCPI, который содержит изображения с разрешением 256x256 пикселей. Это говорит о том, что предлагаемый метод особенно устойчив к OPA-атакам в наборах данных изображений с высоким разрешением.

Дальнейшие исследования включают в себя оценку предложенного подхода с точки зрения защиты от других атак и проведение экспериментов с другими наборами данных.

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

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

Литература

1. Магомадова А. Р., Сапарбиев А. Ш., Натальсон А. В. Влияние искусственного интеллекта на обработку естественного языка в ИТ-технологиях // Научно-технический вестник Поволжья. 2023. № 12. С. 323–325.
2. Зоткина А. А., Шиндина Н. С. Основные задачи NLP и как их решают нейронные сети // Современные информационные технологии. 2023. № 37 (37). С. 14–17.
3. Демьянчук С. В. Современные технологии и интеллектуальные системы в управлении эксплуатацией автотранспорта // Транспортное дело России. 2024. № 1. С. 245–248.
4. Дорофеев Н. А., Снигур Г. Л., Фролов М. Ю., Смирнов А. В., Сасов Д. А., Зубков А. В. Искусственный интеллект, машинное обучение и нейронные сети в морфологии // Вестник Волгоградского государственного медицинского университета. 2024. Т. 21. № 1. С. 3–8. DOI: 10.19163/1994-9480-2024-21-1-3-8.
5. Гарафутдинова Л. В., Каличкин В. К., Федоров Д. С. Объектно ориентированная классификация изображений дистанционного зондирования земли с использованием машинного обучения // Вестник НГАУ (Новосибирский государственный аграрный университет). 2024. № 2 (71). С. 37–47. DOI: 10.31677/2072-6724-2024-71-2-37-47.
6. Карпушкина И. С., Скоморохина Е. Р., Чикенев С. Д. Решение задачи классификации медицинских изображений с помощью методов машинного обучения // Оригинальные исследования. 2023. Т. 13. № 4. С. 79–84.
7. Зоткина А. А. Анализ алгоритмов машинного обучения, используемых в классификации изображений, публикуемых пользователями социальных сетей // Современные информационные технологии. 2023. № 38 (38). С. 38–40.
8. Khan A., Sohail A., Zahoora U., Qureshi A. S. A survey of the recent architectures of deep convolutional neural networks // Artificial Intelligence Review. 2020. Vol. 53, No. 8. Pp. 5455–5516. DOI: 10.1007/s10462-020-09825-6.
9. Rao S., Stutz D., Schiele B. Adversarial training against location optimized adversarial patches // Computer Vision – ECCV 2020 Workshops. ECCV 2020. LNCS. Vol. 12539. 2020. Pp. 429–448. DOI: 10.1007/978-3-030-68238-5_32.
10. Deng Y., Zheng X., Zhang T., Chen C., Lou G., Kim M. An analysis of adversarial attacks and defenses on autonomous driving models // 2020 IEEE International Conference on Pervasive Computing and Communications (PerCom). 2020. P. 1–10. DOI: 10.1109/PerCom45495.2020.9127389.
11. Kotenko I., Saenko I., Lauta O., Vasiliev N., Iatsenko D. Attacks against machine learning systems: analysis and GAN-based approach to protection // Proceedings of the Seventh International Scientific Conference «Intelligent Information Technologies for Industry» (IITI'23). IITI 2023. LNNS. Vol. 777. P. 49–59. 2023. DOI: 10.1007/978-3-031-43792-2_5.
12. Zheng Y., Jie Z., Shiguang S. Adaptive image transformations for transfer-based adversarial attack. ECCV 2022. ECCV 2022. Lecture Notes in Computer Science. 2022. pp. 1–17. DOI: 10.1007/978-3-031-20065-6_1.
13. Zolfi A., Kravchik M., Elovici Y., Shabtai A. The translucent patch: A physical and universal attack on object detectors // 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021. P. 15227–15236. DOI: 10.1109/CVPR46437.2021.01498.
14. Cucu A.-V., Valenzise G., Stănescu D., Ghergulescu I., Găină L. I., Guşiţă B. Defense method against adversarial attacks using JPEG compression and One-Pixel Attack for improved dataset security // 2023 27th International Conference on System Theory, Control and Computing (ICSTCC). 2023. P. 523–527. DOI: 10.1109/ICSTCC59206.2023.10308520.
15. Ning L.-B., Dai Z., Su J., Pan Ch., Wang L., Fan W., Li Q. Interpretation-empowered neural cleanse for backdoor attacks // Companion Proceedings of the ACM Web Conference 2024 (WWW '24). 2024. P. 951–954. DOI: 10.1145/3589335.3651525.
16. Ren K., Zheng T., Qin Zh., Liu X. Adversarial attacks and defenses in deep learning // Engineering. 2020. Vol. 6, No. 3. P. 346–360. DOI: 10.1016/j.eng.2019.12.012.
17. 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.
18. Irfan M. M., Ali S., Yaqoob I., Zafar N. Towards deep learning: A review on adversarial attacks // 2021 International Conference on Artificial Intelligence (ICAI). 2021. P. 91–96. DOI: 10.1109/ICAI52203.2021.9445247.
19. Sharma S., Chen Z. A Systematic study of adversarial attacks against network intrusion detection systems // Electronics. 2024. Vol. 13, No. 24. P. 5030. DOI: 10.3390/electronics13245030.
20. Wang W., Sun J., Wang G. Visualizing One Pixel Attack Using Adversarial Maps // 2020 Chinese Automation Congress (CAC). 2020. P. 924–929. DOI: 10.1109/CAC51589.2020.9327603.

A METHOD OF COUNTERING ADVERSARIAL ATTACKS ON IMAGE CLASSIFICATION SYSTEMS

Kotenko I. V.[11], Saenko I. B.[12], Lauta O. S.[13], Vasiliev N. A.[14], Sadovnikov V. E.[15]

Keywords: *cybersecurity, machine learning, adversarial attacks, image classification, noise pollution, attack defense, artificial intelligence.*

The purpose of the study: *development and evaluation of a method for countering FGSM, ZOO, OPA adversarial attacks on image classification systems based on the integration of noise pollution, neural cleansing and JPEG data compression.*

Research methods: *system analysis, machine learning, image noising, neural cleansing, JPEG data compression, computational experiment.*

Results obtained: *an analysis of works on the topic of attacks on image classification systems (ICS) based on the application of machine learning methods, and methods of protection against them was carried out. Based on the results of this analysis, it was revealed that the most common attacks on ICS include adversarial attacks, namely: Fast Gradient Sign Method (FGSM), Zero-Order Optimization (ZOO) and One Pixel Attack (OPA). The topic of countering these attacks is currently of great interest. The essence of the impact of these attacks on ICS is disclosed, and their influence on the accuracy of image recognition is revealed. A method for countering adversarial attacks is proposed, based on image noising with Gaussian and Poisson noise, as well as the use of JPEG compression and neural cleansing technology. Experiments were conducted showing the high efficiency of the proposed method. The experiments were aimed at assessing the accuracy of image recognition contained in two different data sets - a set of images of personal computer parts and a set of handwritten digital images. The results of image recognition were evaluated before and after exposure of the ICS to adversarial attacks, as well as after applying the proposed method to these sets.*

Scientific novelty: *an analysis of works on the topic of protection against adversarial attacks showed that currently the most typical attacks on ICS are FGSM, ZOO and OPA attacks. The proposed method for countering adversarial attacks on ICS differs from other known protection methods in that it integrates the capabilities of countering attacks contained in three different approaches (neural cleansing, noise pollution and JPEG compression) and identifies the optimal parameters of these approaches. The high efficiency of the proposed method was confirmed in experiments conducted on two different data sets.*

Contribution: *Igor Kotenko and Igor Saenko – general concept of adversarial attacks on ICS and methods of protection against them based on well-known works; Igor Kotenko and Oleg Lauta – description of methods of impact of adversarial attacks; Nikita Vasilev and Vladimir Sadovnikov – implementation of the proposed approach; Igor Kotenko and Igor Saenko – theoretical justification of the proposed approach.*

References

1. Magomadova A. R., Saparbiev A. Sh., Natal'son A. V. Vlijanie iskusstvennogo intellekta na obrabotku estestvennogo jazyka v IT-tehnologijah // Nauchno-tehnicheskij vestnik Povolzh'ja. 2023. № 12. S. 323–325.
2. Zotkina A. A., Shindina N. S. Osnovnye zadachi NLP i kak ih reshajut nejronnye seti // Sovremennye informacionnye tehnologii. 2023. № 37 (37). S. 14–17.
3. Dem'janchuk S. V. Sovremennye tehnologii i intellektual'nye sistemy v upravlenii jekspluataciej avtotransporta // Transportnoe delo Rossii. 2024. № 1. S. 245–248.
4. Dorofeev N. A., Snigur G. L., Frolov M. Ju., Smirnov A. V., Sasov D. A., Zubkov A. V. Iskusstvennyj intellekt, mashinnoe obuchenie i nejronnye seti v morfologii // Vestnik Volgogradskogo gosudarstvennogo medicinskogo universiteta. 2024. T. 21. № 1. S. 3–8. DOI: 10.19163/1994-9480-2024-21-1-3-8.
5. Garafutdinova L. V., Kalichkin V. K., Fedorov D. S. Ob#ektno orientirovannaja klassifikacija izobrazhenij distancionnogo zondirovanija zemli s ispol'zovaniem mashinnogo obuchenija // Vestnik NGAU (Novosibirskij gosudarstvennyj agrarnyj universitet). 2024. № 2 (71). S. 37–47. DOI: 10.31677/2072-6724-2024-71-2-37-47.
6. Karpushkina I. S., Skomorohina E. R., Chikenev S. D. Reshenie zadachi klassifikacii medicinskih izobrazhenij s pomoshh'ju metodov mashinnogo obuchenija // Original'nye issledovanija. 2023. T. 13. № 4. S. 79–84.
7. Zotkina A. A. Analiz algoritmov mashinnogo obuchenija, ispol'zuemyh v klassifikacii izobrazhenij, publikuemyh pol'zovateljami social'nyh setej // Sovremennye informacionnye tehnologii. 2023. № 38 (38). S. 38–40.
8. Khan A., Sohail A., Zahoora U., Qureshi A. S. A survey of the recent architectures of deep convolutional neural networks // Artificial Intelligence Review. 2020. Vol. 53, No. 8. Pp. 5455–5516. DOI: 10.1007/s10462-020-09825-6.

11 Igor V. Kotenko, Honored Worker of Science of the Russian Federation, Dr.Sc. of Technical Sciences, Professor, Chief Scientist and Head of Laboratory of Computer Security Problems at St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg, Russia. E-mail: ivkote@comsec.spb.ru
12 Igor B. Saenko, Dr.S, of Technical Sciences, Professor, Chief Scientist of Laboratory of Computer Security Problems at St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg, Russia. E-mail: ibsaen@comsec.spb.ru
13 Oleg S. Lauta, Dr.Sc. of Technical Sciences, associate professor, Admiral Makarov State University of Maritime and Inland Shipping, St. Petersburg, Russia. E-mail: laos-82@yandex.ru
14 Nikita A. Vasiliev, researcher, Military Telecommunication Academy named after the Soviet Union Marshal Budienny S. M., St. Petersburg, Russia. E-mail: vasn2020@mail.ru
15 Vladimir E. Sadovnikov, graduate student of Laboratory of Computer Security Problems at St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg, Russia. E-mail: bladimir1998@mail.ru

9. Rao S., Stutz D., Schiele B. Adversarial training against location optimized adversarial patches // Computer Vision – ECCV 2020 Workshops. ECCV 2020. LNCS. Vol. 12539. 2020. Pp. 429–448. DOI: 10.1007/978-3-030-68238-5_32.
10. Deng Y., Zheng X., Zhang T., Chen C., Lou G., Kim M. An analysis of adversarial attacks and defenses on autonomous driving models // 2020 IEEE International Conference on Pervasive Computing and Communications (PerCom). 2020. P. 1–10. DOI: 10.1109/PerCom45495.2020.9127389.
11. Kotenko I., Saenko I., Lauta O., Vasiliev N., Iatsenko D. Attacks against machine learning systems: analysis and GAN-based approach to protection // Proceedings of the Seventh International Scientific Conference «Intelligent Information Technologies for Industry» (IITI'23). IITI 2023. LNNS. Vol. 777. P. 49–59. 2023. DOI: 10.1007/978-3-031-43792-2_5.
12. Zheng Y., Jie Z., Shiguang S. Adaptive image transformations for transfer-based adversarial attack. ECCV 2022. ECCV 2022. Lecture Notes in Computer Science. 2022. pp. 1–17. DOI: 10.1007/978-3-031-20065-6_1.
13. Zolfi A., Kravchik M., Elovici Y., Shabtai A. The translucent patch: A physical and universal attack on object detectors // 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021. P. 15227–15236. DOI: 10.1109/CVPR46437.2021.01498.
14. Cucu A.-V., Valenzise G., Stănescu D., Ghergulescu I., Găină L.I., Guşiță B. Defense method against adversarial attacks using JPEG compression and One-Pixel Attack for improved dataset security // 2023 27th International Conference on System Theory, Control and Computing (ICSTCC). 2023. P. 523–527. DOI: 10.1109/ICSTCC59206.2023.10308520.
15. Ning L.-B., Dai Z., Su J., Pan Ch., Wang L., Fan W., Li Q. Interpretation-empowered neural cleanse for backdoor attacks // Companion Proceedings of the ACM Web Conference 2024 (WWW '24). 2024. P. 951–954. DOI: 10.1145/3589335.3651525.
16. Ren K., Zheng T., Qin Zh., Liu X. Adversarial attacks and defenses in deep learning // Engineering. 2020. Vol. 6. P. 346–360. DOI: 10.1016/j.eng.2019.12.012.
17. 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.
18. Irfan M.M., Ali S., Yaqoob I., Zafar N. Towards deep learning: A review on adversarial attacks // 2021 International Conference on Artificial Intelligence (ICAI). 2021. P. 91–96. DOI: 10.1109/ICAI52203.2021.9445247.
19. Sharma S., Chen Z. A Systematic study of adversarial attacks against network intrusion detection systems // Electronics. 2024. Vol. 13, No. 24. P. 5030. DOI: 10.3390/electronics13245030.
20. Wang W., Sun J., Wang G. Visualizing One Pixel Attack Using Adversarial Maps // 2020 Chinese Automation Congress (CAC). 2020. P. 924–929. DOI: 10.1109/CAC51589.2020.9327603.