Автономная некоммерческая образовательная организация высшего образования «Сколковский институт науки и технологий»

На правах рукописи

Паутов Михаил Александрович

Доказуемая устойчивость нейронных сетей

Резюме
диссертации на соискание учёной степени
кандидата компьютерных наук

Научный руководитель:
доктор физико-математических наук, профессор
Оселедец Иван Валерьевич

Москва — 2024

Характеристика работы

Актуальность темы. Алгоритмы на основе нейронных сетей в последнее время достигли огромного успеха в решении практических задач из области компьютерного зрения. В задачах классификации изображений [40], сегментации [41], обнаружения [42] и локализации [43] объектов точность таких алгоритмов зачастую сравнима с точностью человека. Благодаря этому преимуществу они используются в беспилотных автомобилях [44], медицинской диагностике [45], компьютерной безопасности [46] и многих других приложениях. К сожалению, такое преимущество достигается за счет вычислительных затрат на обучение, численной нестабильности и уязвимости к различным видам возмущений входных данных. Некоторое время назад было обнаружено [47], что даже малая окрестность правильно классифицированного изображения заполнена (состязательными) изображениями, которые, хотя и неотличимы от исходного изображения, классифицируются нейронной сетью неправильно. Благодаря этому открытию был сформулирован важный вопрос: как можно проверить правильность предсказания нейронной сети, когда ее входные данные подвергаются преобразованию, не меняющему их семантику (например, добавлению шума ограниченной нормы или повороту изображения на несколько градусов)?

На данный момент в литературе по глубокому обучению предложено множество способов эксплуатации уязвимости нейронных сетей к незначительным изменениям входных данных [48—50; A1]. Как следствие, появилось множество подходов, позволяющих сделать нейронные сети эмпирически более устойчивыми к таким возмущениям [51—53]. Такие методы, хотя и усложняют поиск преобразования входного сигнала, приводящего к неправильной работе нейронной сети, не дают гарантий корректного поведения последней в условии наличия различных преобразований входных данных. Чтобы заполнить этот пробел, появилась новая область теоретического глубокого обучения, названная *доказуемой устойчивостью*. Целью работ в этой области является предоставление гарантий того, что заданная нейронная сеть доказуемо устойчива к определенному типу преобразований входных данных, например, аддитивным возмущениям [54].

В таких приложениях нейронных сетей, как беспилотные автомобили и медицинская диагностика, достоверность предсказаний алгоритмов не менее важна, чем точность. Например, если алгоритм компьютерного зрения обнаруживает пешехода и при дневном свете, и ночью, недостаточно доверить ему управление автомобилем: разработчик должен предоставить *гарантии* того, что пешеход будет обнаружен в разных условиях освещения и погоды.

С практической точки зрения кажется, что обеспечение таких гарантий – неразрешимая задача: требуется гарантировать устойчивость к несчетному множеству преобразованию входных данных. Тем не менее, возможно сформулировать гарантии устойчивости, используя математические свойства нейронной сети.

В области доказуемой устойчивости гарантии корректности поведения нейронных сетей часто обходятся дорогой ценой: либо корректность поведения последних доказывается в очень узком смысле, либо приводит к значительному снижению производительности сертифицированной сети. В связи с этим представляется актуальным разработка методов сертификации к широкому классу входных возмущений, таких, которые не оказывают сильного влияния на производительность моделей. Разработка и интеграция таких методов ведет к расширению спектра практических задач, решение которых можно доверить нейронным сетям.

Степень разработанности темы. Феномен уязвимости нейронных сетей к незначительным (аддитивным) возмущениям во входных данных впервые был описан в работах [47; 55].

В задаче классификации изображений было продемонстрировано, что добавление к правильно классифицируемому нейронной сетью $f : \mathbb{R}^n \to [1, \ldots, K]$ объекту x возмущения, связанного с градиентом используемой при обучении классификатора функции потерь J, часто приводит к неправильной классификации полученного объекта. Именно,

$$f(x) \neq f(x + \delta), \tag{1}$$

где

$$\delta = \epsilon \text{sign}\left(\nabla_x J(\theta, x, y)\right) \tag{2}$$

при небольшом ϵ. Добавки из уравнения (2) получили название состязательных возмущений. В дальнейших работах [48; 50; 56—58] были предложены различные методы построения состязательных возмущений, в том числе, когда потенциальному злоумышленнику недоступно вычисление градиентов нейронной сети. Стоит отметить, что эксплуатация подобных уязвимостей нейронных сетей в физической области является существенно более сложной задачей, но, в то же время, представляющей настоящую угрозу безопасности, например, в таких задачах как распознавание лиц и детекция объектов.

Одни из первых результатов в задаче предоставления гарантий устойчивости классификационных нейронных сетей были описаны в работах [54; 59]. Используемые в данных работах алгоритмы основываются на подходе, называемом случайное сглаживание. Случайное сглаживание заключается в замене исходной модели f ее суррогатом, определенным как

$$g(x) = \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \sigma^2 I)} f(x + \epsilon). \tag{3}$$

В указанных и дальнейших работах показана липшицевость сглаженной функции из уравнения (3) в случае ограниченности скалярной функции f. Данный подход получил развитие в работах [60; 61], где продемонстрирована липшицевость сглаженных моделей по параметру разрешимых преобразований [60].

Среди методов детектирования и противодействия атакам, направленным на кражу функциональности нейронных сетей, наиболее широко представлены подходы, основанные на встраивании цифровых водяных знаков [62]. Важно отметить, что встраивание таких знаков непосредственно в веса нейронной сети обладает серьезным недостатком: часто даже небольшое изменение (весов) модели может привести к потере водяного знака [63]. В последнее время наиболее эффективными методами являются те, что основаны на создании триггерного набора – определенного множества $\mathcal{D}_t = \{(x_i, y_i)\}_{i=1}^n$, объектам которого модель f ставит в соответствие определенные (индикаторные) предсказания. Данные методы работают в предположении о том, что совпадение предсказаний некоторой подозрительной модели f' с предсказаниями модели f на наборе данных $\mathcal{D}_t$ является индикатором того, модель f' есть функциональная копия модели f.

Одним из способов создания триггерного множества может быть наложение текста на изображение с последующим присваиванием полученному объекту метки, отличной от корректной [64]. Важно, что такой подход требует дополнительного обучения исходной модели на новых данных, что, вероятно, может привести к снижению качества модели на исходных данных. Набор триггеров должен быть скрыт от потенциального злоумышленника, поскольку тот может использовать его для того, чтобы добиться от суррогатной модели f' такого же поведения на нем, как и у исходной модели f [65]. Существуют и другие методы подтверждения права собственности, помимо использования цифровых водяных знаков. Существуют подходы, основывающиеся на наблюдении, что расстояние от тренировочных данных до границы принятия решения в среднем больше, чем от тестовых данных: отдельный классификатор может быть обучен для определения принадлежности того или иного объекта к обучающей выборке [66]. В работах [67; 68] состязательные примеры используются в качестве объектов триггерного множества. Стоит отметить, что водяные знаки, нанесенные описанными методами, часто оказываются неустойчивыми к атакам, направленным на кражу функциональности.

Данная работа посвящена проблемам устойчивости и конфиденциальности классификационных нейронных сетей.

Целью данной работы является разработка подходов к обеспечению надежности и конфиденциальности нейронных сетей без заметного снижения производительности последних.

Для достижения поставленной цели необходимо было решить следующие **задачи**:

1. Продемонстрировать уязвимости алгоритмов на основе нейронных сетей в наиболее распространенных практических приложениях.
2. Разработать метод вероятностной сертификации нейронных сетей к возмущениям входных данных произвольного типа.
3. Разработать метод сертификации прототипических нейронных сетей к аддитивным возмущениям ограниченной нормы.
4. Разработать метод на основе водяных знаков для определения кражи нейронной сети, развернутой в условиях “черного ящика”.

Научная новизна:

1. Разработан градиентный метод генерации состязательных патчей для оценки уязвимости моделей распознавания лиц в физической области в реальном времени.
2. Разработан метод вероятностной сертификации нейронных сетей к произвольным возмущениям входных данных.
3. Разработан метод сертификации прототипических нейронных сетей к аддитивным преобразованиям входных данных ограниченной нормы в задаче обучения на малом количестве тренировочных примеров.
4. Разработан метод создания устойчивых водяных знаков для определения кражи нейронной сети, развернутой в условиях “черного ящика”.

Теоретическая и практическая значимость. В работе предложен метод оценки вероятности больших уклонений ограниченной случайной величины на основе модификации неравенства Чернова-Крамера и продемонстрирована липшицевость сглаженных вектор-функций, параметризуемых прототипическими нейронными сетями. Разработаны следующие методы тестирования устойчивости, сертификации и защиты приватности нейронных сетей:

1. Подход к созданию состязательных патчей, приводящих к некорректной работе нейронной сети в задаче распознавания лиц в реальном времени.
2. CC-Cert, метод вероятностной сертификации нейронных сетей к возмущениям входных данных произвольной природы.
3. Smoothed Embeddings, метод сертификации прототипических нейронных сетей к аддитивным преобразованиям входных данных ограниченной нормы в задаче обучения на малом количестве тренировочных примеров.
4. Метод создания устойчивых водяных знаков для определения кражи нейронной сети, развернутой в условиях "черного ящика".

Методология и методы исследования. Результаты диссертации были получены с использованием методов и моделей, применяемых при

сертификации нейронных сетей к преобразованиям входных данных. Математическую основу данной работы составляют теория вероятностей, математическая статистика, линейная алгебра и математический анализ.

Основные положения, выносимые на защиту:

1. Градиентный подход к построению состязательных патчей, демонстрирующих уязвимость систем распознавания лиц в реальном времени в физической области.
2. Вероятностный подход к проверке устойчивости нейронных сетей к входным возмущениям произвольного типа без снижения производительности нейронной сети.
3. Подход к сертификации прототипических нейронных сетей к аддитивным преобразованиям входных данных ограниченной нормы в задаче обучения на малом количестве тренировочных примеров.
4. Метод генерации цифровых водяных знаков в качестве индикатора кражи нейронных сетей без ущерба для производительности исходной модели.

Достоверность результатов диссертации. Научные результаты, описанные в данной работе, являются математическими утверждениями, сопровождаемыми строгими доказательствами. Эффективность представленных в работе методов тестирования устойчивости, сертификации и защиты приватности нейронных сетей подтверждена экспериментально.

Апробация работы. Результаты работы докладывались на конференциях:

1. International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON), Октябрь, 2019. Тема: “On Adversarial Patches: Real-World Attack on ArcFace-100 Face Recognition System”.
2. The Thirty-Sixth AAAI Conference on Artificial Intelligence, Февраль, 2022. Тема: “CC-Cert: A probabilistic approach to certify general robustness of neural networks“.
3. The Thirty-sixth Annual Conference on Neural Information Processing Systems, Декабрь, 2022. Тема: “Smoothed Embeddings for Certified Few-Shot Learning“.
4. Conference Fall into ML, Ноябрь, 2022. Тема: “Smoothed Embeddings for Certified Few-Shot Learning“.
5. ISP RAS Open Conference, Декабрь, 2022. Тема: “Smoothed Embeddings for Certified Few-Shot Learning“.
6. AIRI Seminar AIschnitsa, Декабрь, 2022. Тема: “Smoothed Embeddings for Certified Few-Shot Learning“.
7. The 33rd International Joint Conference on Artificial Intelligence, Чеджу, Южная Корея, Август, 2024. Тема: “Probabilistically Robust Watermarking of Neural Networks“.

Личный вклад. Вклад автора в исследования, описанные в данной диссертации, заключается в следующем:

1. В статье “On Adversarial Patches: Real-World Attack on ArcFace-100 Face Recognition System” [A2] автор реализовал численную процедуру генерации состязательных патчей, приводящих к некорректным предсказаниям системы распознавания лиц в физической области в реальном времени. Автор провел эксперименты по оценке переноса патчей из цифровой области в физическую, а также по интерпретации полученных результатов. Вместе с соавторами автор подготовил текст статьи.
2. В работе “CC-Cert: A probabilistic approach to certify general robustness of neural networks” [A3] автор предложил подход к сертификации нейронных сетей к композициям преобразований в вероятностной постановке. Автором сформулированы и доказаны все теоретические результаты, а именно вероятностные гарантии на корректность предсказания нейронной сети при наличии определенных входных преобразований. Автор также разработал методику оценки предложенного метода, провел эксперименты и интерпретировал их результаты. Совместно с соавторами автор подготовил текст статьи.
3. В работе “Smoothed Embeddings for Certified Few-Shot Learning” [A4] автор предложил подход к сертификации прототипических нейронных сетей к аддитивным возмущениям ограниченной нормы. Автор сформулировал и доказал теоретический результат, а именно детерминированную гарантию правильности предсказания нейронной сети при наличии аддитивных преобразований входного сигнала. Автор также разработал методику оценки предложенного подхода и участвовал в проведении экспериментов. Автор подготовил текст статьи и все его правки.
4. В препринте “Probabilistically Robust Watermarking of Neural Networks” [A5] автор предложил методику генерации надежных цифровых водяных знаков для защиты нейронных сетей от атак, направленных на кражу функциональности. Автор разработал экспериментальную методику для оценки предложенного подхода и вместе с соавторами подготовил текст статьи.

Содержание работы

Во **введении** обосновывается актуальность исследований, проводимых в рамках данной диссертационной работы, формулируется цель, ставятся задачи работы, излагается научная новизна и практическая значимость представляемой работы.

Первая глава является вводной и посвящена демонстрации уязвимости алгоритмов на основе нейронных сетей к аддитивным возмущениям входных данных (состязательным атакам) в физическом домене. В разделах 1.2-1.3 представлена формальная постановка задачи и дано описание некоторых существующих методах создания состязательных атрибутов. Состязательная атака есть возмущение входного объекта нейронной сети, приводящее к некорректному предсказанию на преобразованном объекте. Формально, если задана нейронная сеть $f : \mathbb{R}^n \to \Delta^K$, отображающая входные объекты в векторы вероятностей K классов, порог нормы возмущения ε и входной объект x, то атакой является такое возмущение δ, что

$$\begin{cases} \arg\max_{i\in[1,\dots,K]} f_i(x+\delta) \neq \arg\max_{i\in[1,\dots,K]} f_i(x), \\ \|\delta\| \leq \varepsilon. \end{cases}$$

В разделах 1.4-1.5 приведено описание предложенного градиентного метод построения состязательных возмущений для атаки на нейронную сеть в задаче распознавания лиц. Предложенный метод создания состязательных возмущений основан на итеративном методе решения оптимизационной задачи поиска аддитивной добавки δ для нейронной сети f, отображающей входные объекты в многомерные векторы-эмбеддинги единичной длины, в точке x класса y в цифровом домене:

$$\begin{cases} \gamma^{t+1} = \mu_1\gamma^t + \mu_2 \frac{\nabla_x L(f,x^t,y)}{\|\nabla_x L(f,x^t,y)\|}, \\ \delta^{t+1} = \varepsilon \mathrm{sign}(\gamma^{t+1}), \\ x^{t+1} = \mathrm{clip}_{[0,1]^n}\left[x^t + \delta^{t+1}\right], \\ x^1 = x, \quad \gamma^1 = 0, \quad \delta^0 = 0, \quad \delta = \delta^T. \end{cases}$$

Здесь $L(f, x^t, y)$ есть функция потерь нейронной сети, $\varepsilon, \mu_1, \mu_2$ есть константы.

В работе исследуется переносимость состязательных атрибутов, созданных в цифровом домене, в физический домен. Для обеспечения переносимости состязательных атрибутов в функцию потерь нейронной сети $L(f, x^t, y)$ включены слагаемые, отвечающие за сохранение гладкости создаваемого возмущения в форме TV loss [69]

$$TV(x^t) = \sum_{i,j} \sqrt{\left(x^t_{i,j} - x^t_{i,j+1}\right)^2 + \left(x^t_{i,j} - x^t_{i+1,j}\right)^2}$$

и аугментацию состязательного атрибута преобразованиями из некоторого параметрического множества $\mathcal{T}$ в форме Expectation Over Transformation [70]

$$L_{\mathrm{adv}}(f,x^t,y) = \mathbb{E}_{\tau\in\mathcal{T}} \left(\cos\langle f(\tau(x^t)), c_y\rangle\right),$$

где c_y есть вектор-прототип класса y.

Предложенный метод создания состязательных атрибутов применен в задаче таргетированной состязательной атаки, цель которой заключается в построении возмущения, применение которого приводит к контролируемому изменению предсказания нейронной сети. В работе исследовано влияние расположения состязательного атрибута на эффективность состязательной атаки. В разделе 1.6 приведено описание экспериментов, направленных на демонстрацию эффективности предложенного подхода. В качестве демонстрации эффективности предложенного метода приведены результаты работы системы распознавания лиц после применения состязательных атрибутов (см. Таблицу 1).

Таблица 1 — Численные результаты экспериментов по построению состязательной атаки на систему распознавания лиц LResNet100E-Ir. В качестве метрики используется косинусное подобие между векторами-эмбеддингами классифицируемой фотографии с нанесенным состязательным атрибутом и веторами-прототипами исходного правильного класса (e) и желаемого класса ($e_{x'}$). Объекты из множества x_{train} соответствуют фотографиям, использовавшимся в процессе создания состязательного атрибута; объекты из множества x_{val} соответствуют отложенным фотографиям, использовавшимся для проверки состязательной атаки в цифровом домене; объекты из множества x_{test} соответствуют фотографиям, созданным в физической области с нанесением состязательного атрибута.

Вид состязательного атрибута	Очки	Наклейка на лоб
$cos(e_{x_{train}}, e)$	0.041 ± 0.052	-0.053 ± 0.009
$cos(e_{x_{train}}, e_{x'})$	0.648 ± 0.020	0.221 ± 0.011
$cos(e_{x_{val}}, e)$	0.317 ± 0.004	0.273 ± 0.007
$cos(e_{x_{val}}, e_{x'})$	0.451 ± 0.021	0.421 ± 0.025
$cos(e_{x_{test}}, e)$	0.305 ± 0.024	0.323 ± 0.035
$cos(e_{x_{test}}, e_{x'})$	0.363 ± 0.024	0.391 ± 0.021

Примеры состязательных атрибутов приведены на Рис. 1, 2.

В заключительной части первой главы представлено обсуждение результатов эксперимента и обосновывается важность дальнейших исследований устойчивости нейронных сетей, направленных на создание методов защиты последних от состязательных атак.

Вторая глава посвящена изучению методов предоставления гарантий корректности поведения классификационных нейронных сетей в условиях наличия возмущений входных данных произвольной природы. В этой главе представлено описание предложенного метода предоставления гарантий корректности поведения нейронных сетей в вероятностной постановке. Суть предложенного подхода заключается в оценке вероятности

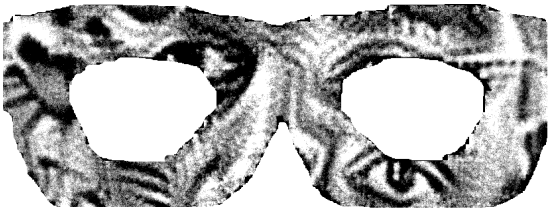


Рис. 1 — Состязательный атрибут в форме оправы очков.

Рис. 2 — Состязательный атрибут в форме наклейки на лоб.

появления ошибочного предсказания нейронной сети в случае, когда исходный правильно классифицируемый объект x подвергается преобразованию $T_\theta(x)$, определенному с точностью до параметров θ, выбирающимися случайным образом из некоторого множества параметров Θ.

В разделе 2.3 работы показано, что если различие между выходными векторами вероятностей $f(x)$ и $f(T_\theta(x))$ по норме L_∞ не превосходит значения $d = \frac{p_1 - p_2}{2}$, где p_1, p_2 есть две максимальные компоненты вектора $f(x)$, то объекты x и $T_\theta(x)$ классифицируются нейронной сетью f как представители одного класса. На основании этого наблюдения и метода Чернова-Крамера оценки вероятности больших отклонений случайной величины предложен подход оценки вероятности неправильной работы нейронной сети. Формально, если $t > 0$, вероятность большого отклонения случайной величины $Z = \|f(x) - f(T_\theta(x)\|_\infty$ ограничена:

$$\mathbb{P}_{\theta \sim \Theta}\left(Z > d\right) \leq e^{-dt}\mathbb{E}(e^{Zt}).$$

Учитывая невозможность вычисления математического ожидания $\mathbb{E}(e^{Zt})$ в общем случае, предлагается метод оценки последнего как функции

$$\hat{b} = \frac{1}{\delta}\max\left\{Y_1, \ldots, Y_k\right\},$$

где $\delta \in (0{,}1)$ есть константа, а $Y_1, \ldots, Y_k$ есть k реализаций выборочного среднего в форме

$$Y_j = \exp(-dt)\frac{1}{n}\sum_{i=1}^{n}\exp(Z_i^j t), \qquad Z_i^j = \|f(x) - f(T_{\theta_i^j}(x)\|_\infty, \qquad \theta_i^j \sim \Theta.$$

Также в разделе 2.3 работы представлены теоретические гарантии применимости предложенного подхода на основе оценки вероятности корректности использования функции от выборочных средних вместо математического ожидания при вычислении вероятности ошибки нейронной сети:

Теорема 1. *Пусть случайная величина Z принимает значения из отрезка* [0,1]*, а плотность распределения случайной величины $\xi = e^{Zt}$ обладает коэффициентом вариации $C_v = \frac{\sigma_\xi}{\mathbb{E}(\xi)}$. Тогда*

$$\mathbb{P}\left(\hat{b} < \frac{\mathbb{E}(\xi)}{e^{dt}}\right) < \left(\frac{1}{1 + \frac{n(1-\delta)^2}{C_v^2}}\right)^k. \tag{4}$$

В разделе 2.4 главы представлено описание экспериментов, направленных на проверку эффективности предложенного метода.

При помощи предложенного метода оценивается вероятность неправильной работы классификационной нейронной сети на публичных наборах данных, а именно MNIST [71] и CIFAR-10 [72].

В качестве результатов экспериментов представлено значение вероятностно сертифицированной точности, *PCA*, в зависимости от порога вероятности ε и было показано, как она связана с эмпирически сертифицированной точностью, *ERA*. Для заданного классификатора $h(\cdot) = \arg\max_{k\in\{1,\dots,K\}} f(\cdot)$, набора изображений $\mathcal{S} = \{(x_1, y_1), \dots, (x_m, y_m)\}$, функции вычисления вероятности больших уклонений $\hat{b} = \hat{b}(x)$ и порога ε, вероятностно сертифицированная точность вычисляется как

$$PCA(\mathcal{S}, \varepsilon) = \frac{\left|(x, y) \in \mathcal{S} : \hat{b}(x) < \varepsilon \;\&\; h(x) = y\right|}{m}. \tag{5}$$

Для разбиения $\Theta = \{\theta_1, \dots, \theta_r\}$ пространства параметров преобразования T эмпирически сертифицированная точность вычисляется как доля объектов из $\mathcal{S}$, которые правильно классифицируются нейронной сетью после применения любого из преобразований $T_{\theta_i}, i \in [1, \dots, r]$:

$$ERA(\mathcal{S}) = \frac{|(x, y) \in \mathcal{S} : h(T_{\theta_i}(x)) = y \;\forall i \in [1, \dots r]|}{m}. \tag{6}$$

Также в разделе 2.4 приведено описание параметров экспериментов и рассмотренных преобразований входных данных. Результаты экспериментов представлены в Таблице 2. В Таблице 3 приведено описание параметров рассмотренных в рамках экспериментов преобразований. Стоит отметить, что некоторые параметры зависят от характеристик исходных изображений (например, от размера).

Таблица 2 — Сравнение вероятностно сертифицированной точности и эмпирически сертифицированой точности. В столбце Преобразование использована следующая нотация: B соответствует преобразованию яркости, C – контраста, R – поворота, G – гауссова размытия, T – переноса, S – масштабирования. В столбце Обучение P означает обычное обучение модели, A – обучение с аугментацией данных. Приведены значения метрики PCA для трех порогов ε: высокой ($\varepsilon < 10^{-10}$), средней ($\varepsilon < 10^{-7}$) и низкой ($\varepsilon < 10^{-4}$) степени уверенности в сертификации. В столбце PA указана точность модели на отложенной выборке.

Набор данных	**Преобразование**	**Обучение**	**ERA**	**PCA(ε)** $\varepsilon = 10^{-10}$	$\varepsilon = 10^{-7}$	$\varepsilon = 10^{-4}$	**PA**
CIFAR-10	B	P	58.4%	47.8%	51.6%	55.2%	91.18%
		A	65.0%	55.4%	59.4%	61.8%	88.67%
	C	P	91.6%	62.4%	67.0%	69.6%	91.18%
		A	88.0%	67.0%	72.8%	74.2%	88.67%
	R	P	73.4%	64.6%	69.0%	71.0%	91.18%
		A	72.4%	57.4%	63.6%	67.4%	87.77%
	G	P	12.2%	11.0%	11.0%	11.0%	91.18%
		A	60.4%	57.2%	57.2%	57.8%	81.11%
	T	P	40.4%	28.0%	31.2%	35.2%	91.18%
		A	35.0%	17.8%	22.4%	25.6%	85.98%
	S	P	57.0%	54.4%	54.4%	54.4%	91.18%
		A	55.0%	53.4%	53.4%	53.6%	86.76%
	C + B	P	0.0%	0.0%	0.0%	0.0%	91.18%
		A	0.4%	0.0%	0.0%	0.0%	88.67%
	R + B	P	22.6%	16.2%	20.6%	21.8%	91.18%
		A	30.4%	21.2%	24.6%	27.6%	84.50%
	S + B	P	10.2%	10.4%	10.4%	10.4%	91.18%
		A	41.8%	40.6%	40.6%	40.6%	86.53%
MNIST	B	P	97.8%	94.8%	96.4%	97.0%	99.26%
		A	98.6%	97.0%	98.2%	98.2%	99.04%
	C	P	98.8%	96.0%	97.0%	97.2%	99.26%
		A	98.6%	98.2%	98.2%	98.2%	99.04%
	R	P	18.8%	11.6%	14.8%	16.4%	99.26%
		A	98.0%	97.0%	97.4%	97.6%	99.01%
	G	P	78.0%	68.8%	68.8%	68.8%	99.26%
		A	97.8%	97.8%	97.8%	97.8%	98.35%
	T	P	0.0%	0.0%	0.0%	0.0%	99.26%
		A	39.6%	31.4%	34.4%	38.2%	99.09%
	S	P	21.6%	21.0%	21.0%	21.0%	99.26%
		A	34.4%	34.4%	34.4%	34.4%	99.25%
	C + B	P	8.4%	0.0%	0.0%	0.0%	99.26%
		A	7.6%	2.4%	2.4%	2.4%	99.04%
	R + B	P	14.0%	9.2%	11.2%	13.0%	99.26%
		A	95.2%	93.0%	93.4%	94.6%	99.08%
	S + B	P	13.0%	13.4%	13.4%	13.4%	99.26%
		A	93.4%	93.0%	93.0%	93.4%	99.37%

Таблица 3 — Параметры преобразований, рассмотренных в рамках экспериментов.

Набор данных	**Преобразование**	**Значение параметров**
CIFAR-10	Яркость	$\theta_b \in [-40\%, 40\%]$
	Контраст	$\theta_c \in [-40\%, 40\%]$
	Поворот	$\theta_r \in [-10^\circ, 10^\circ]$
	Гауссово размытие	$\theta_g \in [0, 3]$ – радиус ядра
	Перенос	$\lvert\theta_t\rvert \leq 20\%$
	Масштабирование	$\theta_s \in [70\%, 130\%]$
	Контраст + Яркость	см. Контраст & Яркость
	Поворот + Яркость	см. Поворот & Яркость
	Масштабирование + Яркость	см. Масштабирование & Яркость
MNIST	Яркость	$\theta_b \in [-50\%, 50\%]$
	Контраст	$\theta_c \in [-50\%, 50\%]$
	Поворот	$\theta_r \in [-50^\circ, 50^\circ]$
	Гауссово размытие	$\theta_g \in [0, 3]$ – радиус ядра
	Перенос	$\lvert\theta_t\rvert \leq 30\%$
	Масштабирование	$\theta_s \in [70\%, 130\%]$
	Контраст + Яркость	см. Контраст & Яркость
	Поворот + Яркость	см. Поворот & Яркость
	Масштабирование + Яркость	см. Масштабирование & Яркость

В заключительной части главы приведено обсуждение применимости представленного подхода для сертификации устойчивости нейронной сети к случайному преобразованию входного сигнала и определено одно из перспективных направлений дальнейших исследований – анализ устойчивости классификационных нейронный сетей к преобразованиям входных данных произвольной природы и предоставление детерминированных гарантий устойчивости.

Третья глава посвящена исследованию устойчивости прототипических нейронных сетей к аддитивным возмущениям входных данных ограниченной нормы. В вводном разделе 3.1 приведена общая постановка задачи создания алгоритмов, обладающих доказуемой устойчивостью к аддитивным возмущениям входных данных, и мотивация к созданию подобных систем в контексте решения задач классификации в условиях обучения нейронной сети на малом количестве тренировочных примеров.

В разделе 3.2 приведена формальная постановка задачи: для заданной нейронной сети

$$f : \mathbb{R}^D \to \mathbb{R}^d,$$

ставящей в соответствие исходным объектам единичные вектора-эмбеддинги размерности d, набора данных $\mathcal{S} = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, где $x_i \in \mathbb{R}^D$ – классифицируемые объекты, а $y_i \in \{1, \ldots, K\}$, и классификационного правила $h(x) = \arg\min_{k \in \{1,\ldots,K\}} \|f(x) - c_k\|_2$, требуется построить классификатор g, доказуемо устойчивый к аддитивным возмущениям δ малой нормы, или, другими словами, построить такой классификатор g, что равенство

$$h(x) = h(x + \delta)$$

выполняется для всех $\delta : \|\delta\| \leq t$ для некоторого порога t. Здесь

$$c_k = \frac{1}{S_k} \sum_{x \in S_k} f(x)$$

есть вектор-прототип класса k, вычисленный как средний вектор-эмбеддинг объектов класса k (обозначенных за S_k).

В разделе 3.3 представлено описание подхода к созданию классификаторов, доказуемо устойчивых к аддитивным возмущениям входных данных ограниченной нормы, основывающийся на идее применения случайного сглаживания [54]. В основе метода – замена исходной прототипической нейронной сети $f : \mathbb{R}^D \to \mathbb{R}^d$ на суррогатную нейронную сеть, определенную как

$$g(x) = \mathbb{E}_{\varepsilon \sim \mathcal{N}(0, \sigma^2 I)} f(x + \varepsilon).$$

Для сглаженной прототипической нейронной сети в форме $g(x)$ в разделе 3.3 сформулировано и доказано свойство липшицевости в форме теоремы.

Теорема 2. *(Константа Липшица) Предположим, что $f : \mathbb{R}^D \to \mathbb{R}^d$ – детерминированная функция и $g(x) = \mathbb{E}_{\varepsilon \sim \mathcal{N}(0, \sigma^2 I)} f(x + \varepsilon)$ непрерывно дифференцируема при всех x. Если для всех x $\|f(x)\|_2 = 1$, то $g(x)$ является $L-$липшицевой по l_2-норме с константой $L = \sqrt{\frac{2}{\pi\sigma^2}}$.*

Для доказуемо устойчивой классификации в пространстве эмбеддингов оценено максимальное отклонение эмбеддинга классифицируемого объекта, которое не изменяет ближайший прототип класса. Выражение для данного расстояния сформулировано в форме теоремы.

Теорема 3. *(Состязательный риск) Пусть дано исходное изображение $x \in \mathbb{R}^D$ и сглаженная прототипическая нейронная сеть $g : \mathbb{R}^D \to \mathbb{R}^d$.*

Тогда ближайшая точка к границе разделения классов в пространстве эмбеддингов находится на расстоянии:

$$\gamma = \|\Delta\|_2 = \frac{\|c_2 - g(x)\|_2^2 - \|c_1 - g(x)\|_2^2}{2\|c_2 - c_1\|_2^2}, \tag{7}$$

где $c_1 \in \mathbb{R}^d$ и $c_2 \in \mathbb{R}^d$ – два ближайших прототипа классов. Значение γ – это расстояние между эмбеддингом классифицируемого объекта x и границей разделения классов, представленных прототипами c_1 и c_2. Заметим, что это минимальное по l_2-норме возмущение в пространстве эмбеддингов, необходимое для изменения предсказания g.

С учетом свойств липшицевости и величины максимального отклонения в пространстве эмбеддингов Δ, не изменяющего результат классификации, сформулирована гарантия устойчивости прототипической нейронной сети к аддитивным возмущениям:

Теорема 4. *(Гарантия устойчивости) Нижняя граница l_2-нормы возмущения, изменяющего предсказание классификатора g из теоремы 2 имеет вид*

$$r = \frac{\gamma}{L}, \tag{8}$$

где L - константа Липшица из теоремы 2, а γ – состязательный риск из теоремы 3. Значение r – это сертифицированный радиус g в точке x, или, другими словами, минимальное по l_2-норме возмущение во входном пространстве, необходимое для изменения предсказания g.

В разделе 3.4 приведено описание процедуры сертификации. В разделе 3.5 представлено описание экспериментов и приведены их результаты.

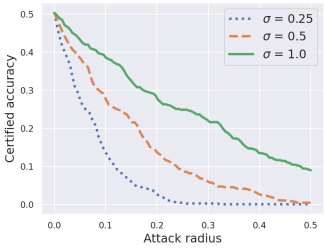
Для оценки предложенного подхода используется несколько наборов данных для обучения исходной прототипической сети. *Cub-200-2011* [73] – это набор данных с 11788 изображений 200 видов птиц, где 5864 изображений 100 видов находятся в обучающем подмножестве, а 5924 изображений других 100 видов – в тестовом подмножестве. *mini*ImageNet [74] – это подмножество изображений из набора данных *ILSVRC 2015* [75] с 64 категориями изображений в обучающем подмножестве, 16 категориями в валидационном подмножестве и 20 категориями в тестовом подмножестве с 600 изображениями размером 84×84 в каждой категории. *CIFAR FS* [76] является подмножеством набора данных *CIFAR 100* [72], содержащим 37800 изображений из 64 категорий в обучающем наборе и 11400 изображений из 20 категорий в тестовом наборе.

Для оценки эффективности предложенного подхода приведены значения сертифицированной точности алгоритма, CRA. Для данного набора объектов S, сглаженной прототипической сети g с соответствующим

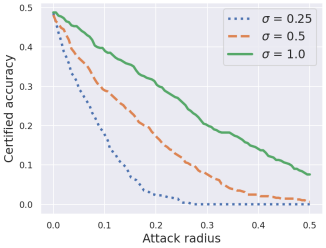
классификационным правилом $h(x) = \arg\min_{i\in\{1,\dots,K\}} \|g(x) - c_k\|_2$, ограничения ε на l_2-норму аддитивного возмущения и функции вычисления гарантии $r = r(x)$ из теоремы 4, метрика CRA определяется как:

$$CRA(S,\varepsilon) = \frac{|(x,y) \in S : r(x) > \varepsilon \;\&\; h(x) = y|}{|S|}. \tag{9}$$

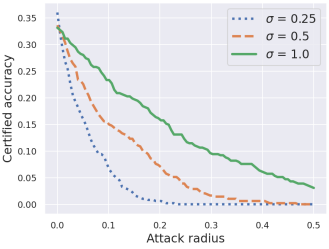
На рисунках 3-4 представлена зависимость сертифицированной точности от величины максимальной нормы аддитивного возмущения для различных сценариев обучения базовой модели (для случаев 1 и 5 обучающих примеров на класс). Значение переменной “Attack radius” соответствует ограничению ε из уравнения (9).



Cub-200-2011



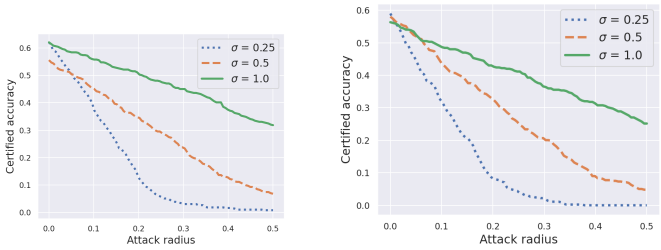
CIFAR-FS



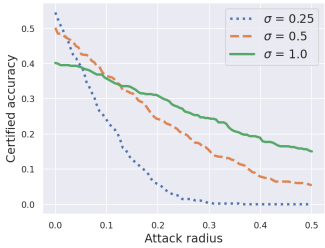
*mini*ImageNet

Рис. 3 — Зависимость сертифицированной точности от порога ε для различных значений σ, случай 1 тренировочного объекта на класс.

В разделе 3.6 приведена теоретическая оценка границ применимости предложенного метода. В заключительном разделе 3.7 приведено обсуждение результатов экспериментов и сформулированы возможные направления дальнейших исследований – обобщение предложенного подхода на другие типы возмущений входных данных и снижение вычислительной сложности процедуры сертификации прототипических нейронных сетей.



Cub-200-2011 CIFAR-FS



*mini*ImageNet

Рис. 4 — Зависимость сертифицированной точности от порога ε для различных значений σ, случай 5 тренировочных объектов на класс.

Четвертая глава посвящена созданию цифровых водяных знаков в качестве индикаторов атак, направленных на кражу функциональности нейронной сети.

В разделе 4.1 приведена мотивация к разработке методов создания водяных знаков, устойчивых к атакам, направленным на кражу функциональности нейронной сети. В разделе отмечается необходимость защиты конфиденциальности нейронных сетей, развернутых по принципу "черного ящика": потенциальный злоумышленник может прибегнуть к дистилляции нейронной сети, получив таким образом функционально похожую модель, не тратя время и ресурсы на разработку и обучение нейронной сети. В разделах 4.2-4.3 представлен обзор актуальной литературы и представлена необходимая справочная информация об атаках, направленных на кражу функциональности моделей, и методах создания цифровых водяных знаков.

В разделе 4.4 приведено описание предложенного подхода к созданию водяных знаков на основе триггерного набора, устойчивых к кражам функциональности нейронной сети.

Под триггерным набором в работе понимается множество $\mathcal{D}_t^* \subset \mathbb{R}^n$ такое, что предсказания исходной нейронной сети $f : \mathbb{R}^n \to \Delta^k$ на объектах из него являются заранее определенными. Предлагаемая в работе процедура создания триггерного набора состоит из двух частей:

– В первой части происходит набор точек-кандидатов на включение в триггерный набор. Пусть исходная нейронная сеть f обучена на наборе данных $\mathcal{D}$ и дан отложенный набор данных $\mathcal{D}_h : \mathcal{D}_h \cap \mathcal{D} = \emptyset$. Путь пара точек $(x_{i_1}, y_{i_1}), (x_{i_2}, y_{i_2})$ из разных классов $y_{i_1} \neq y_{i_2}$ выбрана случайно и равновероятно из $\mathcal{D}_h$. Тогда кандидат на включение в триггерных набор имеет вид

$$x_i^* = \lambda x_{i_1} + (1 - \lambda) x_{i_2},$$

где $\lambda \sim \mathcal{U}(0,1)$. Чтобы гарантировать нестандартное заранее определенное поведение модели f на объекте x_i^*, он принимается в качестве кандидата только в том случае, если исходная модель определяет x_i^* как представителя некоторого другого класса y_i^* : $y_i^* \neq y_{i_1}, y_i^* \neq y_{i_2}$.

– Во второй части происходит верификация кандидатов на включение в триггерный набор. Отметим, что совпадение предсказаний исходной модели f и некоторой подозрительной модели f^* на объектах из триггерного набора должно быть индикатором того, что модели f^* есть функциональная копия модели f. Для удовлетворения этого свойства вводится параметрический набор прокси-моделей $\mathcal{B}_{\delta,\tau}(f)$, который имитирует множество подозрительных моделей. Множество прокси-моделей определяется двумя параметрами, δ и τ, регулирующими отличие прокси-моделей от исходной модели в пространстве весов и в смысле точности решения исходной задачи. Формально оно задается в виде

$$\mathcal{B}_{\delta,\tau}(f) = \{f' : \|\theta(f') - \theta(f)\|_2 \leq \delta \text{ и } |\text{acc}(\mathcal{D}, f') - \text{acc}(\mathcal{D}, f)| \leq \tau\},$$

где $\theta(f)$ есть вектор весов модели f, а $\text{acc}(\mathcal{D}, f)$ есть точность модели f на наборе данных $\mathcal{D}$.

Далее для верификации переносимости поведения на триггерном наборе случайно выбираются m прокси-моделей $f_1, \ldots, f_m \in \mathcal{B}_{\delta,\tau}(f)$. Затем проверяется, все ли m прокси-моделей присваивают объекту из триггерного набора ту же метку класса, что и исходная модель f. Иными словами, объект-кандидат (x_i^*, y_i^*) включается в триггерный набор, если выполняется условие

$$y_i^* = f(x_i^*) = f_1(x_i^*) = \cdots = f_m(x_i^*).$$

В разделе 4.5 приведено описание экспериментов. В качестве обучающих выборок используются наборы данных CIFAR-10 и CIFAR-100

[72], в качестве исходной модели используется сверточная нейронная сеть ResNet34 [40]. Предложенный метод сравнивается с существующими подходами к созданию цифровых водяных знаков в задаче детектирования факта кражи модели при помощи дистилляции в следующих постановках:

– Атака типа Soft-label. В этом случае выход $f(x)$ исходной модели представляет собой вектор вероятностей классов. Суррогатная модель f^* обучается путем минимизации функционала

$$L_{\text{ext}}(\hat{\mathcal{D}}) = \frac{1}{|\hat{\mathcal{D}}|} \sum_{\hat{x}_i \in \hat{\mathcal{D}}} D_{KL}(f(\hat{x}_i), f^*(\hat{x}_i)),$$

где D_{KL} есть дивергенция Кульбака-Лейблера, а $\hat{\mathcal{D}}$ есть отложенная выборка данных, используемая для копирования функциональности модели.

– Атака типа Hard-label. В этом случае выход $f(x)$ исходной модели – это метка класса, присвоенная моделью f объекту x. Такая постановка соответствует обучению суррогатной модели на наборе данных $\hat{\mathcal{D}} = \{x_i, f(x_i)\}_{i=1}^N$.

– Атака с регуляризацией. В работе [63] было предложено обучение суррогатной модели путем одновременной минимизации ошибки последней на обучающем наборе данных $\mathcal{D}$ и KL-дивергенции между выходами исходной модели и суррогатной модели.

В качестве результатов экспериментов приводятся значения точности исходной и суррогатных моделей на тестовом подмножестве набора данных $\mathcal{D}$, а также точность моделей на триггерном наборе данных $\mathcal{D}_t^*$. Результаты экспериментов представлены в Таблице 4.

В разделе 4.6 главы приведено обсуждение ограничений предложенного метода. В заключительном разделе главы приведено обсуждение результатов экспериментов и определено одно из направлений дальнейших исследований – разработка гарантий переносимости предсказания на триггерных наборах на суррогатные модели.

Метод	Метрика	f	Суррогатные модели f^*		
			Soft-label	Hard-label	RGT
EWE [77]	CIFAR-10	86.10 ± 0.54	83.97 ± 1.02	82.22 ± 0.50	88.88 ± 0.35
RS [78]		84.17 ± 1.01	88.93 ± 1.18	89.62 ± 0.97	90.14 ± 0.08
MB [63]	acc. (%)	87.81 ± 0.76	91.17 ± 0.76	91.88 ± 0.40	93.05 ± 0.20
Ours		$\mathbf{91.00 \pm 0.00}$	$\mathbf{92.60 \pm 0.91}$	$\mathbf{94.87 \pm 0.59}$	$\mathbf{99.42 \pm 0.02}$
EWE [77]	Триггерный набор	26.88 ± 8.32	51.01 ± 5.58	36.05 ± 6.48	1.64 ± 1.05
RS [78]		95.67 ± 4.93	7.67 ± 4.04	6.33 ± 1.15	3.00 ± 0.00
MB [63]	acc. (%)	100.00 ± 0.00	82.00 ± 1.00	51.33 ± 4.93	72.67 ± 6.66
Ours		100.00 ± 0.00	$\mathbf{85.10 \pm 6.33}$	$\mathbf{73.70 \pm 4.65}$	$\mathbf{78.00 \pm 5.58}$
EWE [77]	CIFAR-100	55.11 ± 1.67	53.00 ± 1.57	46.78 ± 1.00	63.73 ± 0.40
RS [78]		59.87 ± 2.78	65.66 ± 1.53	65.79 ± 0.39	64.99 ± 0.30
MB [63]	acc. (%)	62.13 ± 4.36	$\mathbf{67.66 \pm 0.36}$	$\mathbf{70.65 \pm 0.49}$	$\mathbf{70.24 \pm 0.46}$
Ours		$\mathbf{66.70 \pm 0.00}$	67.49 ± 0.03	68.05 ± 0.73	67.85 ± 0.04
EWE [77]	Триггерный набор	68.14 ± 10.16	30.90 ± 11.34	15.10 ± 5.64	5.73 ± 3.42
RS [78]		99.00 ± 1.00	2.67 ± 1.53	4.33 ± 4.16	2.00 ± 1.00
MB [63]	acc. (%)	100.00 ± 0.00	70.67 ± 7.57	40.00 ± 8.89	62.66 ± 10.12
Ours		100.00 ± 0.00	$\mathbf{78.80 \pm 2.93}$	$\mathbf{74.70 \pm 3.16}$	$\mathbf{79.10 \pm 2.77}$

Таблица 4 — Показатели эффективности методов нанесения водяных знаков для разных методов атак, направленных на кражу функциональности. Лучшие показатели выделены жирным шрифтом. Стоит отметить, что предложенный метод позволяет создавать триггерные наборы, наиболее эффективные в качестве индикаторов кражи функциональности нейронной сети.

В **заключении** кратко формулируются основные результаты работы. Они заключаются в реализации предложенных методов оценки устойчивости и приватности нейронных сетей:

- Градиентный метод создания состязательных атрибутов, демонстрирующий уязвимость классификационных нейронных сетей в практических приложениях (на примере задачи распознавания лиц);
- Метод предоставления гарантий устойчивости нейронных сетей к возмущениям входных данных произвольной природы в вероятностной постановке;
- Метод предоставления гарантий устойчивости прототипических нейронных сетей к аддитивным возмущениям входных данных ограниченной нормы;
- Метод создания цифровых водяных знаков, устойчивых к атакам, направленным на кражу функциональности нейронных сетей.

Публикации автора по теме диссертации

A1. Real-World Attack on MTCNN Face Detection System [Текст] / E. Kaziakhmedov [и др.] // 2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON). — IEEE. 2019. — С. 0422—0427.

A2. On Adversarial Patches: Real-World Attack on ArcFace-100 Face Recognition System [Текст] / M. Pautov [и др.] // 2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON). — IEEE. 2019. — С. 0391—0396.

A3. CC-Cert: A probabilistic Approach to Certify General Robustness of Neural Networks [Текст] / M. Pautov [и др.] // Proceedings of the AAAI Conference on Artificial Intelligence. Т. 36. — 2022. — С. 7975—7983.

A4. Smoothed Embeddings for Certified Few-Shot Learning [Текст] / M. Pautov [и др.] // Advances in Neural Information Processing Systems. — 2022. — Т. 35. — С. 24367—24379.

A5. Probabilistically Robust Watermarking of Neural Networks [Текст] / M. Pautov [и др.] // arXiv preprint arXiv:2401.08261. — 2024.

Список литературы

1. Deep residual learning for image recognition [Текст] / K. He [и др.] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. — 2016. — С. 770—778.
2. Image segmentation using deep learning: A survey [Текст] / S. Minaee [и др.] // IEEE transactions on pattern analysis and machine intelligence. — 2021. — Т. 44, № 7. — С. 3523—3542.
3. Object detection in 20 years: A survey [Текст] / Z. Zou [и др.] // Proceedings of the IEEE. — 2023.
4. Efficient object localization using convolutional networks [Текст] / J. Tompson [и др.] // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2015. — С. 648—656.
5. End to end learning for self-driving cars [Текст] / M. Bojarski [и др.] // arXiv preprint arXiv:1604.07316. — 2016.
6. *Sarvamangala, D.* Convolutional neural networks in medical image understanding: a survey [Текст] / D. Sarvamangala, R. V. Kulkarni // Evolutionary intelligence. — 2022. — Т. 15, № 1. — С. 1—22.
7. *Aydin, I.* A new IoT combined face detection of people by using computer vision for security application [Текст] / I. Aydin, N. A. Othman // 2017 International Artificial Intelligence and Data Processing Symposium (IDAP). — IEEE. 2017. — С. 1—6.
8. Intriguing properties of neural networks [Текст] / C. Szegedy [и др.] // arXiv preprint arXiv:1312.6199. — 2013.

9. *Moosavi-Dezfooli, S.-M.* Deepfool: a simple and accurate method to fool deep neural networks [Текст] / S.-M. Moosavi-Dezfooli, A. Fawzi, P. Frossard // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. — 2016. — С. 2574—2582.

10. Universal adversarial perturbations [Текст] / S.-M. Moosavi-Dezfooli [и др.] // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2017. — С. 1765—1773.

11. *Carlini, N.* Towards evaluating the robustness of neural networks [Текст] / N. Carlini, D. Wagner // 2017 ieee symposium on security and privacy (sp). — IEEE. 2017. — С. 39—57.

12. Adversarial defense via learning to generate diverse attacks [Текст] / Y. Jang [и др.] // Proceedings of the IEEE/CVF International Conference on Computer Vision. — 2019. — С. 2740—2749.

13. Towards deep learning models resistant to adversarial attacks [Текст] / A. Madry [и др.] // arXiv preprint arXiv:1706.06083. — 2017.

14. *Zhou, J.* Manifold Projection for Adversarial Defense on Face Recognition [Текст] / J. Zhou, C. Liang, J. Chen // European Conference on Computer Vision. — Springer. 2020. — С. 288—305.

15. *Cohen, J.* Certified adversarial robustness via randomized smoothing [Текст] / J. Cohen, E. Rosenfeld, Z. Kolter // International Conference on Machine Learning. — PMLR. 2019. — С. 1310—1320.

16. *Goodfellow, I. J.* Explaining and Harnessing Adversarial Examples [Текст] / I. J. Goodfellow, J. Shlens, C. Szegedy // CoRR. — 2014. — Т. abs/1412.6572.

17. *Chen, J.* Hopskipjumpattack: A query-efficient decision-based attack [Текст] / J. Chen, M. I. Jordan, M. J. Wainwright // 2020 IEEE Symposium on Security and Privacy (SP). — IEEE. 2020. — С. 1277—1294.

18. Black-box adversarial attacks with limited queries and information [Текст] / A. Ilyas [и др.] // International conference on machine learning. — PMLR. 2018. — С. 2137—2146.

19. Geoda: a geometric framework for black-box adversarial attacks [Текст] / A. Rahmati [и др.] // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — 2020. — С. 8446—8455.

20. Certified robustness to adversarial examples with differential privacy [Текст] / M. Lecuyer [и др.] // 2019 IEEE Symposium on Security and Privacy (SP). — IEEE. 2019. — С. 656—672.

21. Tss: Transformation-specific smoothing for robustness certification [Текст] / L. Li [и др.] // Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. — 2021. — С. 535—557.

22. Gsmooth: Certified robustness against semantic transformations via generalized randomized smoothing [Текст] / Z. Hao [и др.] // International Conference on Machine Learning. — PMLR. 2022. — С. 8465—8483.

23. *Hartung, F.* Multimedia watermarking techniques [Текст] / F. Hartung, M. Kutter // Proceedings of the IEEE. — 1999. — Т. 87, № 7. — С. 1079—1107.

24. Margin-based neural network watermarking [Текст] / B. Kim [и др.] // International Conference on Machine Learning. — PMLR. 2023. — С. 16696—16711.

25. Protecting intellectual property of deep neural networks with watermarking [Текст] / J. Zhang [и др.] // Proceedings of the 2018 on Asia conference on computer and communications security. — 2018. — С. 159—172.

26. Evaluating the robustness of trigger set-based watermarks embedded in deep neural networks [Текст] / S. Lee [и др.] // IEEE Transactions on Dependable and Secure Computing. — 2022. — Т. 20, № 4. — С. 3434—3448.

27. *Maini, P.* Dataset Inference: Ownership Resolution in Machine Learning [Текст] / P. Maini, M. Yaghini, N. Papernot // International Conference on Learning Representations.

28. *Le Merrer, E.* Adversarial frontier stitching for remote neural network watermarking [Текст] / E. Le Merrer, P. Perez, G. Trédan // Neural Computing and Applications. — 2020. — Т. 32, № 13. — С. 9233—9244.

29. *Lukas, N.* Deep Neural Network Fingerprinting by Conferrable Adversarial Examples [Текст] / N. Lukas, Y. Zhang, F. Kerschbaum // International Conference on Learning Representations.

30. *Rudin, L. I.* Nonlinear total variation based noise removal algorithms [Текст] / L. I. Rudin, S. Osher, E. Fatemi // Physica D: nonlinear phenomena. — 1992. — Т. 60, № 1—4. — С. 259—268.

31. Synthesizing Robust Adversarial Examples [Текст] / A. Athalye [и др.] // ICML. — 2017.

32. *Deng, L.* The mnist database of handwritten digit images for machine learning research [best of the web] [Текст] / L. Deng // IEEE signal processing magazine. — 2012. — Т. 29, № 6. — С. 141—142.

33. Learning multiple layers of features from tiny images [Текст] / A. Krizhevsky, G. Hinton [и др.]. — 2009.

34. The caltech-ucsd birds-200-2011 dataset [Текст] / C. Wah [и др.]. — 2011.

35. Matching networks for one shot learning [Текст] / O. Vinyals [и др.] // Advances in neural information processing systems. — 2016. — Т. 29. — С. 3630—3638.

36. Imagenet large scale visual recognition challenge [Текст] / O. Russakovsky [и др.] // International journal of computer vision. — 2015. — Т. 115, № 3. — С. 211—252.

37. Meta-learning with differentiable closed-form solvers [Текст] / L. Bertinetto [и др.] // arXiv preprint arXiv:1805.08136. — 2018.

38. Entangled watermarks as a defense against model extraction [Текст] / H. Jia [и др.] // 30th USENIX Security Symposium (USENIX Security 21). — 2021. — С. 1937—1954.

39. Certified neural network watermarks with randomized smoothing [Текст] / A. Bansal [и др.] // International Conference on Machine Learning. — PMLR. 2022. — С. 1450—1465.