Теоретические основы информатики

Theoretical foundations of computer science

УДК 004.9

УСТОЙЧИВОСТЬ НЕЙРОННЫХ СЕТЕЙ К СОСТЯЗАТЕЛЬНЫМ АТАКАМ ПРИ РАСПОЗНАВАНИИ БИОМЕДИЦИНСКИХ ИЗОБРАЖЕНИЙ

Д. М. ВОЙНОВ[1)]***, В. А. КОВАЛЕВ***[2)]

[1)]*Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Беларусь*
[2)]*Объединенный институт проблем информатики НАН Беларуси, ул. Сурганова, 6, 220012, г. Минск, Беларусь*

В настоящий момент большинство исследований и разработок в области глубокого обучения концентрируются на повышении точности распознавания, в то время как проблема состязательных атак на глубокие нейронные сети и их последствий пока не получила должного внимания. Данная статья посвящена экспериментальной оценке влияния различных факторов на устойчивость нейронных сетей к состязательным атакам при решении задач распознавания биомедицинских изображений. На обширном материале, включающем более чем 1,45 млн радиологических и гистологических изображений, исследуется эффективность атак, подготовленных с помощью алгоритма спроецированного градиентного спуска (PGD), алгоритма «глубокого обмана» (DeepFool) и алгоритма Карлини – Вагнера (CW). Анализируются результаты атак обоих типов (по методам белого и черного ящика) на

Образец цитирования:
Войнов ДМ, Ковалев ВА. Устойчивость нейронных сетей к состязательным атакам при распознавании биомедицинских изображений. *Журнал Белорусского государственного университета. Математика. Информатика.* 2020;3:60–72.
https://doi.org/10.33581/2520-6508-2020-3-60-72

For citation:
Voynov DM, Kovalev VA. The stability of neural networks under condition of adversarial attacks to biomedical image classification. *Journal of the Belarusian State University. Mathematics and Informatics.* 2020;3:60–72. Russian.
https://doi.org/10.33581/2520-6508-2020-3-60-72

Авторы:
Дмитрий Михайлович Войнов – магистрант кафедры дискретной математики и алгоритмики факультета прикладной математики и информатики. Научный руководитель – В. А. Ковалев.
Василий Алексеевич Ковалев – кандидат технических наук; заведующий лабораторией анализа биомедицинских изображений.

Authors:
Dmitry M. Voynov, master's degree student at the department of discrete mathematics and algorithmics, faculty of applied mathematics and computer science.
voynovdd@gmail.com
Vassili A. Kovalev, PhD (engineering); head of the laboratory of biomedical image analysis.
vassili.kovalev@gmail.com

нейронные сети с архитектурами InceptionV3, Densenet121, ResNet50, MobileNet и Xception. Основной вывод работы заключается в том, что проблема состязательных атак актуальна для задач распознавания биомедицинских изображений, поскольку протестированные алгоритмы успешно атакуют обученные нейронные сети так, что их точность падает ниже 15 %. Установлено, что при тех же величинах злонамеренных возмущений изображения алгоритм PGD менее эффективен, чем алгоритмы DeepFool и CW. При использовании в качестве метрики сравнения изображений L_2-нормы алгоритмы DeepFool и CW генерируют атакующие изображения близкого качества. В трех из четырех задач распознавания радиологических и гистологических изображений атаки по методу черного ящика с использованием алгоритма PGD показали низкую эффективность.

Ключевые слова: глубокое обучение; состязательные атаки; биомедицинские изображения.

THE STABILITY OF NEURAL NETWORKS UNDER CONDITION OF ADVERSARIAL ATTACKS TO BIOMEDICAL IMAGE CLASSIFICATION

D. M. VOYNOV [a], ***V. A. KOVALEV*** [b]

[a]*Belarusian State University, 4 Niezaliežnasci Avenue, Minsk 220030, Belarus*
[b]*United Institute of Informatics Problems, National Academy of Sciences of Belarus, 6 Surhanava Street, Minsk 220012, Belarus*

Corresponding author: V. A. Kovalev (vassili.kovalev@gmail.com)

Recently, the majority of research and development teams working in the field deep learning are concentrated on the improvement of the classification accuracy and related measures of the quality of image classification whereas the problem of adversarial attacks to deep neural networks attracts much less attention. This article is dedicated to an experimental study of the influence of various factors on the stability of convolutional neural networks under the condition of adversarial attacks to biomedical image classification. On a very extensive dataset consisted of more than 1.45 million of radiological as well as histological images we assess the efficiency of attacks performed using the projected gradient descent (PGD), DeepFool and Carlini – Wagner (CW) methods. We analyze the results of both white and black box attacks to the commonly used neural architectures such as InceptionV3, Densenet121, ResNet50, MobileNet and Xception. The basic conclusion of this study is that in the field of biomedical image classification the problem of adversarial attack stays sharp because the methods of attacks being tested are successfully attacking the above-mentioned networks so that depending on the specific task their original classification accuracy falls down from 83–97 % down to the accuracy score of 15 %. Also, it was found that under similar conditions the PGD method is less successful in adversarial attacks comparing to the DeepFool and CW methods. When the original images and adversarial examples are compared using the L_2-norm, the DeepFool and CW methods generate the adversarial examples of similar maliciousness. In addition, in three out of four of black-box attacks, the PGD method has demonstrated lower attacking efficiency.

Keywords: deep learning; adversarial attacks; biomedical images.

Введение

Одним из активно развивающихся и повсеместно применяемых инструментов современного машинного обучения является глубокое обучение – использование глубоких нейронных сетей в качестве обучаемого алгоритма. Глубокие нейронные сети показывают высокие результаты в широком спектре задач машинного обучения, таких как анализ изображений (классификация, сегментация, обнаружение (или детектирование) объектов), анализ текста (определение содержания, выделение смысла), обработка звука (распознавание речи) и др. Причинами этого можно назвать, во-первых, их способность выявлять сложнейшие зависимости в данных, а во-вторых, колоссальную «вместимость», что позволяет как ученым, так и инженерам не задумываясь выбирать нейронные сети, если размер данных очень велик.

На сегодняшний день большинство разрабатываемых методов и алгоритмов направлены на достижение максимальной точности работы обучаемых моделей [1]. Однако такой подход привел научное сообщество к серьезной проблеме. Обнаружилось, что глубокие нейронные сети чрезвычайно неустойчивы. Были найдены специальные алгоритмы, которые минимально изменяют входное изображение, после чего оно по каким-то причинам неправильно распознается сетью [2]. При этом изменения изображения настолько малы, что зачастую неразличимы человеческим глазом. Процесс, при котором такое

изображение генерируется и подается на вход сети, называется состязательной атакой. Разумеется, указанная проблема создает серьезную брешь в безопасности нейронных сетей и ставит под сомнение целесообразность их использования в задачах с высокой ответственностью. Поэтому крайне важным является изучение и устранение данного эффекта.

В настоящей работе исследуется такая область применения глубоких нейронных сетей, как анализ биомедицинских изображений. Использование машинного обучения в этой области необходимо для решения множества задач [3; 4], на основе которых разрабатываются так называемые системы автоматического диагностирования. На последних лежит колоссальная ответственность, поскольку от их работы может зависеть человеческая жизнь. Кроме того, как и любая другая предметная область, анализ биомедицинских изображений, помимо общих задач и характеристик, имеет свою специфику. Поэтому проведение исследований на примере задач из указанной области не только помогает изучать общую проблему атак на нейронные сети как таковую, но и позволяет получить новые экспериментальные данные непосредственно по этому классу приложений.

Состязательные атаки

Понятие состязательных атак. Под состязательной атакой понимается процесс, в результате которого атакуемый классификатор предсказывает класс изображения неверно как с точки зрения человека, так и с точки зрения обученной и протестированной нейросетевой модели. Под ошибкой предсказания понимается тот факт, что изображения, которые классификатор идентифицирует неправильно, являются, на первый взгляд, вполне допустимыми для соответствующей предметной области, но относятся к другому классу. При этом разница между такими ошибками и простыми ошибками распознавания заключается в том, что для данных изображений обычно есть практически не отличающиеся от них парные, которые тем не менее распознаются сетью правильно.

В настоящей работе исследование состязательных атак будет проводиться на примере задач классификации (распознавания) биомедицинских изображений. Однако стоит отметить, что атаки такого рода можно проводить и в рамках других задач машинного обучения (например, в области анализа и распознавания звука).

Генерация атакующих изображений. Первым и до сих пор основным способом проведения состязательных атак является генерация атакующих изображений (*adversarial examples*), т. е. искусственных изображений, слабо отличающихся от «нормальных» изображений определенной предметной области. Атакующие изображения генерируются с помощью специальных алгоритмов. Благодаря построению данных алгоритмов и некоторым другим факторам (о них будет сказано позже) подача таких изображений на вход нейронной сети зачастую заканчивается ошибкой предсказания.

Генерация атакующих изображений заключается в последовательном выполнении трех шагов:

- выбора изображения из предметной области атакуемой нейронной сети;
- генерации особого шумоподобного возмущения при помощи специального алгоритма;
- применения полученного возмущения к выбранному изображению путем обычного попиксельного сложения.

В результате получается искомое атакующее изображение, которое, вероятно, будет ошибочно предсказано классификационной сетью. Дадим формальное определение атакующим изображениям.

Атакующие изображения

Определение атакующего изображения. Пусть $x \in \mathbb{R}^d$ – нормализованное входное изображение; $y : \mathbb{R}^d \to (0, 1)^p$ – выход классификационной нейронной сети как функции от входного изображения с количеством классов p; $F : (0, 1)^p \to \{1, \ldots, p\}$ – решающая функция классификации (в данной работе рассматривается функция argmax). Положим $\varepsilon > 0$ – некоторое небольшое положительное число. Тогда ε-атакующим изображением называется такое изображение, для которого справедливо неравенство

$$F\left(y\left(x\right)\right) \neq F\left(y\left(x^*\right)\right) \tag{1}$$

при выполнении ограничения

$$\left\|x - x^*\right\| < \varepsilon. \tag{2}$$

В последнем уравнении в качестве нормы рассматривают, как правило, L_2 или L_∞, но, разумеется, допустима любая норма.

Уравнение (1) показывает, что результат классификации атакующего изображения отличен от такового у исходного изображения. Но это обычная ситуация, например, для изображений, принадлежащих

другому классу. Для того чтобы продемонстрировать неестественность эффекта, вводится ограничение (2) на малую величину ε. Приемлемое значение параметра (такое, чтобы атакующее изображение было достаточно близко к исходному) выбирается атакующим субъектом. Обычно, чтобы заставить классификатор ошибиться, достаточно выбрать небольшое ε. Этот параметр называется магнитудой модификации, поскольку условие (2) можно переписать в виде

$$\|\Delta x\| < \varepsilon,\ x^* = x + \Delta x.$$

В таком случае ε является магнитудой модификации изображения x. Разумеется, при разных значениях ε модификация изображения выражена в различной степени. Многие работы показывают, что часто изображение можно изменить незаметным для глаза человека образом и все равно заставить классификатор ошибиться.

Алгоритмы генерации атакующих изображений

На сегодняшний день разработано достаточно много алгоритмов генерации атакующих изображений. Среди них выделяются как универсальные концепции, так и множество разнообразных эвристик. Для начала рассмотрим общую идею существующих алгоритмов, а затем приведем несколько конкретных методов.

Классификация алгоритмов генерации атакующих изображений. По информации, необходимой для работы, алгоритмы генерации делят на атаки по методу белого ящика и атаки по методу черного ящика.

Для проведения атаки по методу белого ящика необходимо знать конфигурацию сети, включая ее архитектуру и все параметры, полученные в результате обучения. Кроме того, требуется наличие оригинального изображения соответствующей предметной области для генерации самого́ атакующего изображения.

Для проведения атаки по методу черного ящика достаточно иметь доступ ко входу сети, куда подаются изображения, и к результатам предсказания, в то время как конфигурация сети может оставаться неизвестной. Разумеется, также нужна информация о предметной области распознаваемых изображений.

По наличию или отсутствию предопределенного класса, который необходимо фальсифицировать атакующему изображению, алгоритмы генерации атак делятся на направленные и ненаправленные. При проведении направленной атаки класс, к которому должно быть ошибочно отнесено атакующее изображение, заранее определен. Ненаправленные атаки, в свою очередь, призваны лишь обмануть сеть вне зависимости от того, к какому ошибочному классу будет отнесено изображение в результате распознавания. Большинство алгоритмов генерации атакующих изображений позволяют сконфигурировать их для проведения атаки любого из двух указанных типов. В данной работе рассматриваются исключительно ненаправленные атаки. Очевидно также, что в случае бинарной классификации типа «норма/патология» направленные и ненаправленные атаки совпадают.

Алгоритм спроецированного градиентного спуска (*projected gradient descent*, PGD). Данный метод есть не что иное, как применение классической техники градиентного спуска с учетом ограниченности возмущения [5–7]. В качестве целевой функции можно рассмотреть k-ю компоненту вектора вероятностей y'. Тогда минимизация такой функции будет приводить к снижению вероятности принадлежности атакующего изображения этому классу, а максимизация – к повышению.

В результате генерация направленного на класс t атакующего изображения по этому методу зависит от коэффициента обучения $\alpha > 0$, количества итераций $n \in \mathbb{N}$, магнитуды возмущений ε и определяется следующим образом:

$$x_{k+1} = \mathrm{clip}_{x,\varepsilon}\left(x_k + \alpha \nabla y_t(x_k)\right),$$

где $x_0 = x$, $k \in [0,\ n-1]$, $x^* = x_n$, а функция $\mathrm{clip}_{x,\varepsilon}$ «обрезает» те элементы ее аргумента, которые отличаются от тех же элементов x более чем на ε. Генерация ненаправленного атакующего изображения по этому методу зависит от исходного класса m и определяется следующим образом:

$$x_{k+1} = \mathrm{clip}_{x,\varepsilon}\left(x_k - \alpha \nabla y_m(x_k)\right).$$

Поскольку на каждой итерации применяется функция $\mathrm{clip}_{x,\varepsilon}$, то в результате работы алгоритма получится изображение x^*, автоматически удовлетворяющее ограничению (2) для L_∞-нормы.

Алгоритм «глубокого обмана» (DeepFool). Данный алгоритм основан на идее линеаризации функции выхода нейронной сети и итеративном вычислении атакующего изображения как точки проекции на некоторую псевдополуплоскость [8; 9]. Одна итерация этого алгоритма задается формулой

$$x_{k+1} = x_k - \frac{y_l(x_k) - y_m(x_k)}{\|w_l - w_m\|^2}(w_l - w_m),$$

где $w = \nabla y(x_k)$, $x_0 = x$; m – исходный класс объекта x, а l – класс, выбираемый на каждой итерации так, чтобы возмущения объекта были минимальны. В качестве атакующего изображения используется значение $x^* = (1+\eta)x_p$ той итерации, на которой атака оказалась успешной.

Алгоритм Карлини – Вагнера (CW). Данный алгоритм генерации атакующих изображений основан на модификации применения алгоритма L-BFGS к специально поставленной задаче оптимизации [10]. Для генерации направленных на класс t атакующих изображений авторами работы [11] формулируется следующая задача минимизации:

$$\left\|\frac{1}{2}(\tanh(w)+1) - x\right\| + cf_t\left(\frac{1}{2}(\tanh(w)+1)\right) \to \min,\ w \in \mathbb{R},$$

где c – конфигурируемый параметр, а функция $f_t(x)$ задается формулой

$$f_t(x) = \max\left(\max_{k \neq t}(y_k(x)) - y_t(x),\ -\kappa\right).$$

Функция f_t также зависит от параметра κ, показывающего желаемую степень уверенности в классификации атакующего изображения. При κ = 0 достаточно найти успешное атакующее изображение, вероятность принадлежности целевому классу которого просто больше вероятностей принадлежности другим классам, но при увеличении κ повышается. Далее эта функция минимизируется известным оптимизатором Adam, и в результате после ограниченного количества итераций получается атакующее изображение.

Используемые наборы данных

Исходные наборы изображений. В проведенном исследовании использовались пять различных наборов биомедицинских изображений, представленных в таблице. Ниже дается их краткое описание.

Задачи классификации и точность их решения
Image classification tasks and their classification accuracy

Набор данных	Акроним	Задача классификации	Количество изображений		Точность
			Общее	По классам	
Гистология с метастазами	H-MT	Норма/участки с метастазами	100 000	50 000/50 000	0,97
Гистология с опухолями в яичниках или щитовидной железе	H-OV	В яичниках: норма/опухоль	96 000	48 000/48 000	0,92
	H-TH	В щитовидной железе: норма/опухоль	96 000	48 000/48 000	0,94
	H-OV-TH	Яичники – норма/ яичники – опухоль/ щитовидная железа – норма/ щитовидная железа – опухоль	192 000	48 000/48 000/ 48 000/48 000	0,91
Рентген легких (норма)	X-NR2	Две возрастные группы: 20–35/50–70 лет	200 000	100 000/100 000	0,98
	X-NR3	Три возрастные группы: 17–24/25–41/42–80 лет	550 080	183 360/183 360/ 183 360	0,83
Компьютерная томография легких	CT	Норма/туберкулез	149 248	111 990/37 258	0,96
Гистологические изображения, окрашенные шестью химическими агентами	H-ST	Препараты: CD31/CD105/D240/ FRES/H&E/Ki67	267 984	59 568/37 488/ 55 296/35 280/ 24 192/56 160	0,95

Гистологические изображения тканей лимфоузлов, пораженных метастазами. Первоначально имелся набор полнослайдовых цветных гистологических изображений, окрашенных с использованием широко распространенной методики гематоксилин – эозин. Размеры изображений достигали 100 000 × 100 000 пк. Поскольку изображения такого размера обычно не подвергаются анализу целиком, они были разрезаны на непересекающиеся плитки размером 256 × 256 пк. Полученные плитки были очищены от участков чистого стекла и артефактов, которые не несут никакой информации. В итоге был сформирован набор из 100 000 цветных изображений размером 256 × 256 пк, представляющих два класса – норму и участки с метастазами. Набор был полностью сбалансирован по классам: каждый из них включал 50 000 изображений.

Гистологические изображения тканей яичников и щитовидной железы, пораженных опухолями. Оригинальный набор данных состоял из 4000 изображений биопсии 26 пациентов. Изображения размером 2048 × 1536 пк представляли либо норму, либо участки злокачественных опухолей в каждом из упомянутых органов. Применив технику, аналогичную описанной выше, мы получили набор из 192 000 цветных изображений размером 256 × 256 пк, разделенных на четыре класса: яичники – норма, яичники – опухоль, щитовидная железа – норма, щитовидная железа – опухоль. Данный набор также был сбалансирован: каждый класс включал 48 000 изображений.

Рентгеновские изображения легких. Оригинальный набор содержал около 2 млн (точно – 1 908 926) рентгеновских изображений грудной клетки разного размера, снятых с помощью различных цифровых рентгеновских аппаратов в процессе скрининга населения в целях выявления заболеваний легких, сердечно-сосудистой системы и скелета. Каждое изображение сопровождалось текстовым отчетом врача-радиолога, который включал первичный диагноз, а также информацию о поле и возрасте пациента. Изображения были нормализованы по яркости и приведены к одинаковому размеру (512 × 512 пк) путем применения передовых алгоритмов интерполяции. Из полученного массива данных было выбрано подмножество, состоящее из более 0,5 млн (точно – 550 080) изображений грудной клетки мужчин и женщин в возрасте от 17 до 80 лет включительно. Отбор осуществлялся с условием обеспечения равномерного представительства пациентов по полу и возрастным группам.

Компьютерная томография легких. Первоначально набор данных состоял из множества трехмерных компьютерно-томографических изображений пациентов, больных туберкулезом легких. Ввиду большой размерности 3D-изображения были разбиты на аксиальные 2D-слои размером 512 × 512 пк. Затем в целях сохранения пространственной информации каждое полученное полутоновое двумерное изображение было преобразовано в цветное путем размещения изображения текущего слоя в зеленый канал, последующего нижнего слоя – в красный канал, а предыдущего верхнего слоя – в синий. Иными словами, пачка из трех слоев, отображающая вариабельность структуры 3D-изображения по оси z, представлялась в виде псевдоцветного изображения с тремя каналами – R, G и B. Поскольку размерность этих изображений все еще оставалась большой, они были нарезаны на плитки размером 256 × 256 пк. В результате получилось 149 248 цветных изображений размером 256 × 256 пк. Описанный набор данных был несбалансирован, так как включал 111 990 изображений со здоровыми участками легких и 37 258 изображений, представляющих легкие с новообразованиями различных видов, вызванных туберкулезом (каверны, фокусы, плеврит и др.).

Гистологические изображения тканей, окрашенные шестью гистохимическими препаратами. Оригинальный набор данных состоял из полнослайдовых гистологических изображений тканей, окрашенных шестью различными иммуногистохимическими препаратами (антителами), широко используемыми при диагностике онкологических заболеваний. Аналогично вышеописанным гистологическим наборам данных эти изображения также были нарезаны на плитки размером 256 × 256 пк. В результате был сформирован набор из 267 984 цветных изображений, разделенных на шесть классов. Данные о балансе классов приведены в таблице.

Таким образом, исследовалось большое количество биомедицинских изображений различных модальностей, которые являются весьма распространенными и часто используемыми в медицине. В частности, гистологические изображения служат золотым стандартом в диагностике рака мягких тканей, а рентгеновские и компьютерно-томографические – одним из основных инструментов при диагностике заболеваний легких и выявлении дефектов скелета.

Вычислительные эксперименты

Построенные задачи классификации. На основе пяти вышеприведенных наборов биомедицинских изображений были построены восемь задач классификации (см. таблицу). Для некоторых наборов данных существовала возможность конфигурирования нескольких задач классификации, что позволило изучить эффект состязательных атак более детально. Как видно из таблицы, в настоящей работе

рассматриваются в основном задачи бинарной классификации, которые наиболее часто встречаются на практике, либо те, к которым обычно сводятся более сложные задачи дифференциальной диагностики заболеваний.

Обучение нейронных сетей. Для каждой задачи классификации была обучена сверточная нейронная сеть. В качестве архитектуры сети выбрана InceptionV3. Предобученные веса сверток не использовались. Для обучения сетей изображения нормировались до отрезка [0, 1]. Тренировочным оптимизатором выступал Adam с одинаковым для всех задач обучающим коэффициентом. Во всех случаях для достижения приемлемых для исследования точностей классификации потребовалось менее 50 эпох обучения. Вычисления проводились на компьютере с процессором Intel® Core™ i7-6700K и двумя видеокартами Nvidia GeForce GTX 1080 Ti. В качестве библиотек для обучения нейронных сетей использовались Keras и Tensorflow.

Исследование атак по методу белого ящика

Теперь опишем проводимые эксперименты. Для каждой пары *задача классификации – алгоритм генерации* выполняются следующие действия.

1. На нейронную сеть, обученную для решения задачи классификации, для каждого изображения из тестовой выборки при помощи выбранного алгоритма проводится атака. В результате генерируется атакующее изображение (рис. 1). Для дальнейшей оценки качества алгоритмов DeepFool и CW L_2- и L_∞-нормы разности сгенерированного и исходного изображений сохраняются, а алгоритм PGD запускается для нескольких значений ε.

2. Атакующее изображение подается на вход этой же сети (атака по методу белого ящика), а полученные вероятности принадлежности классам сохраняются для дальнейшего анализа.

В целях оценки эффективности работы алгоритма и качества генерируемых атакующих изображений вычисляется доля успешных атак, в результате которых норма разности атакующего и исходного изображений ограничена некоторым числом. В настоящей работе рассматривались L_2- и L_∞-нормы.

Зависимость успешности атак от L_∞-нормы возмущения. Построим график зависимости доли успешных атак от L_∞-нормы применяемого возмущения. Как было сказано выше, для возможности построить такую зависимость алгоритм PGD запускается отдельно для ε, равного 0,02–0,20 с шагом 0,02 (соответственно, напрямую ограничивая L_∞-норму), а для алгоритмов DeepFool и CW L_∞-нормы возмущения вычисляются по завершении их работы. Результаты представлены на рис. 2.

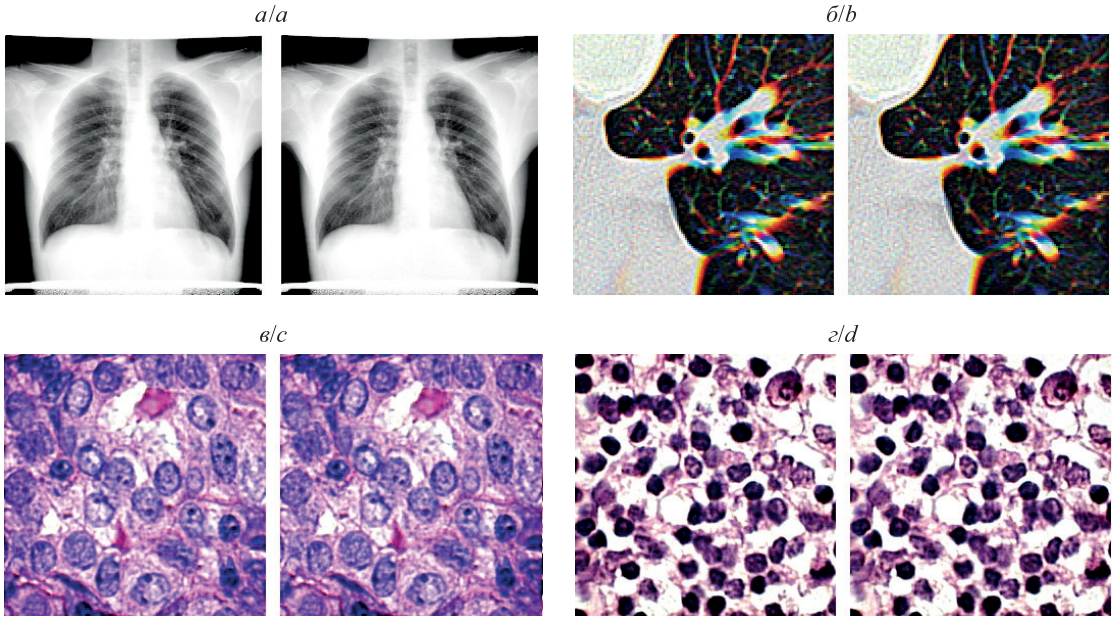


Рис. 1. Примеры радиологических (верхний ряд) и гистологических (нижний ряд) изображений. В каждой паре слева показано исходное изображение, а справа – его атакующая версия, полученная с помощью алгоритма CW

Fig. 1. Examples of radiological (top row) and histological (bottom row) images. In each image pair the original image is given on the left whereas its adversarial version is presented on the right side

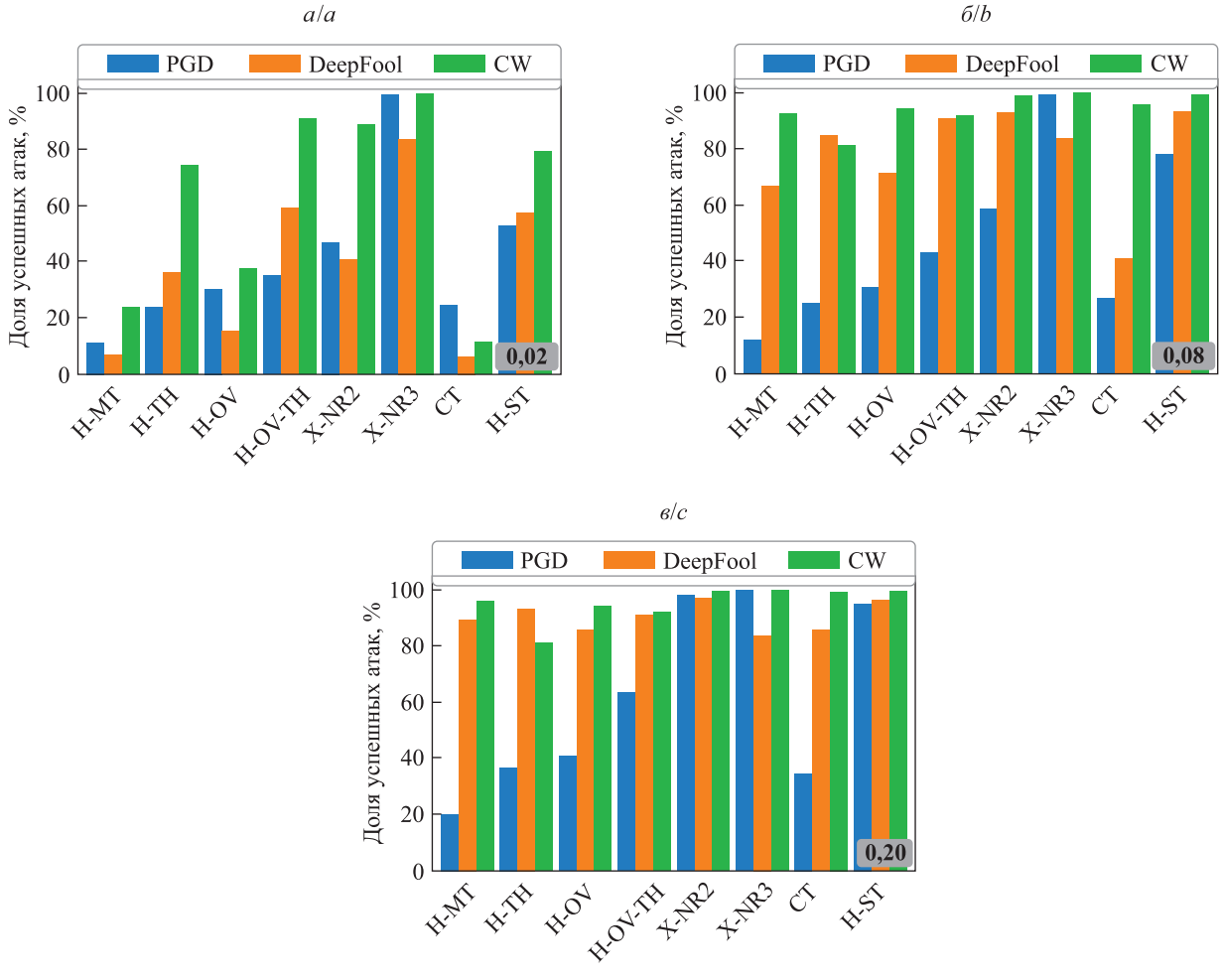


Рис. 2. Доля успешных атак для использованных наборов данных при ограничении L_∞-нормы возмущения в 0,02; 0,08 и 0,20

Fig. 2. The fraction of successful attacks for each classification task under condition of limit of L_∞ perturbations equal to 0.02; 0.08 and 0.20

При использовании L_∞-нормы алгоритм CW показал себя лучше почти во всех случаях (кроме ε, равного 0,08 и 0,20 на наборе данных H-TH, где эффективнее себя проявил алгоритм DeepFool). Также можно отметить, что при минимальном рассмотренном значении ε алгоритм PGD часто немного более эффективен, чем DeepFool. Однако уже при ε не ниже 0,08 качество работы последнего заметно превышает результат PGD. В целом, как и ожидалось, алгоритм PGD показывал себя хуже, чем DeepFool и CW, несмотря на то что последние оптимизируют L_2-норму. При этом значительной эффективности (хотя бы 80 %) PGD не достигает в шести из восьми случаев даже при максимальном анализируемом возмущении.

Зависимость успешности атак от L_2-нормы возмущения. Построим аналогичную описанной выше зависимость, ограничивая возмущение L_2-нормой. Данную зависимость будем строить для алгоритмов DeepFool и CW, поскольку, как было показано ранее, алгоритм PGD работает заметно хуже при средних и больших возмущениях, а запуская его только для маленького ε, мы не получим высокой итоговой эффективности. Результаты приведены на рис. 3.

Из представленных данных видно, что в целом алгоритмы DeepFool и CW близки по эффективности: для ε, равного 1,0 и 2,0, в четырех из восьми случаев доли успешных атак этих алгоритмов почти равны, в оставшихся случаях однозначного фаворита не наблюдается. Однако стоит отметить, что для изображений размером 256 × 256 пк ε, равный 2,0, для L_2-нормы возмущения является весьма небольшой величиной. При указанном значении каждый пиксел в среднем меняется на 0,007 8 (в условиях нормировки [0, 1]), что составляет менее 1 % от максимально допустимого значения.

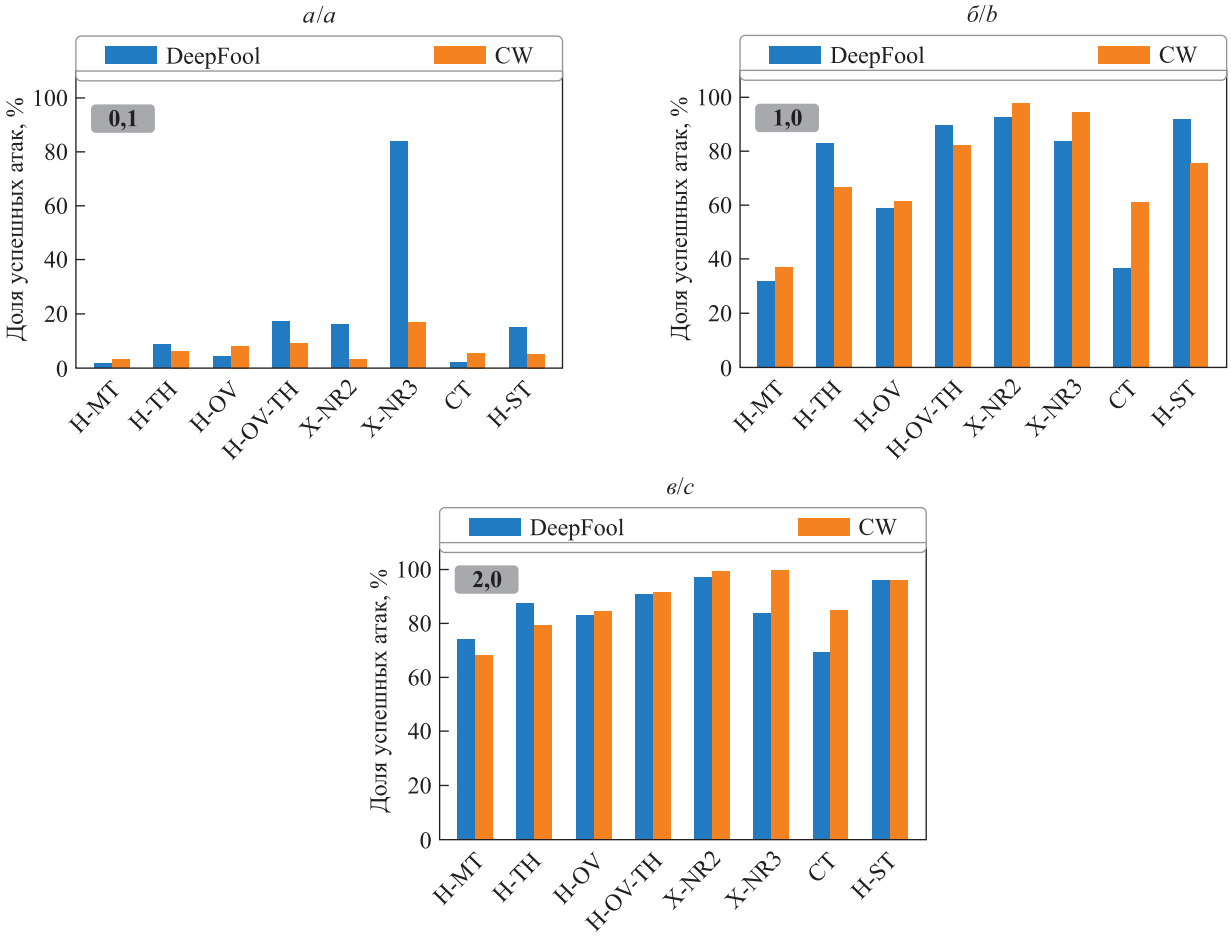


Рис. 3. Доля успешных атак для восьми наборов данных при ограничении L_2-нормы возмущения в 0,1; 1,0 и 2,0

Fig. 3. The fraction of successful attacks for each classification task under condition of limit of L_2 perturbations equal to 0.1; 1.0 and 2.0

Исследование атак по методу черного ящика

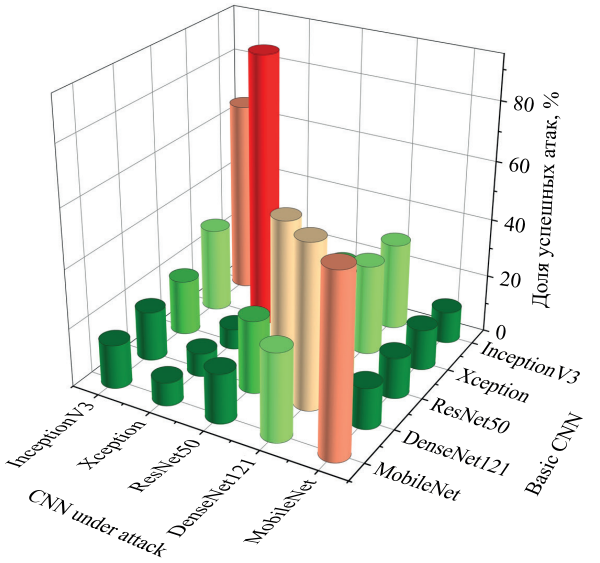
Методика проведения атак. Атаки по методу черного ящика осуществляются без использования информации об архитектуре или обученных весах нейронной сети. Нетрудно заметить, что в таких условиях генерация атакующих изображений по представленной выше методике становится невозможной, поскольку все рассмотренные алгоритмы используют градиент функции выхода нейронной сети, который напрямую зависит от обученных параметров. В данном случае требуется принципиально новая методика проведения атак.

На сегодняшний день предложено несколько способов осуществления атак по методу черного ящика. Большинство из них основаны на свойстве переносимости, заключающемся в том, что атакующее изображение, сгенерированное для атаки одной сети, часто успешно атакует и другую сеть, обученную классифицировать изображения того же типа [12]. Несмотря на то что в настоящий момент теоретические обоснования такого явления отсутствуют, оно неоднократно наблюдалось на практике. Полагаясь на это свойство, можно сформулировать методику действий, состоящую из трех основных шагов:

- на некоторой выборке изображений из предметной области классификации целевой сети обучаем свою инструментальную (имитирующую) сеть (в данной работе рассматривается режим, при котором доступна тренировочная выборка изображений атакуемой целевой сети);
- проводим атаку по методу белого ящика на обученную инструментальную сеть и в итоге получаем атакующее изображение;
- подаем сгенерированное атакующее изображение на вход целевой атакуемой сети и оцениваем получаемые результаты.

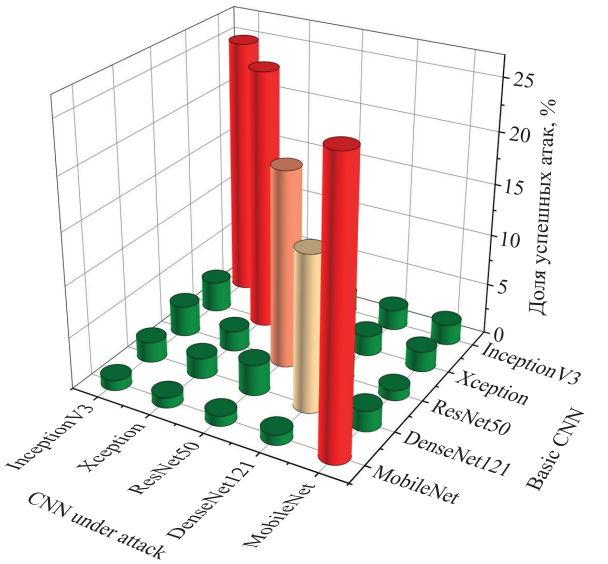
Таким образом, целевая сеть атакуется по методу черного ящика, поскольку информация об архитектуре и весах целевой сети никак не использовалась, потребовалась только информация о соответствующих параметрах обученной инструментальной сети, играющей роль вспомогательного инструмента.

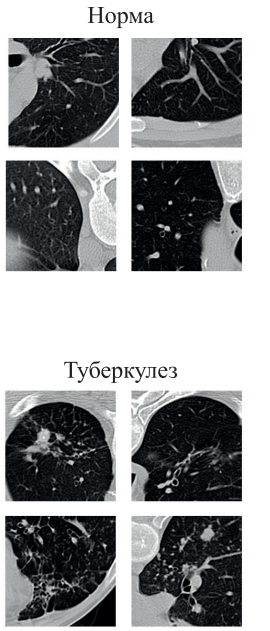
а/*a*



CNN under attack	Basic CNN				
	InceptionV3	Xception	ResNet50	DenseNet121	MobileNet
InceptionV3	66	29	19	17	15
Xception	13	93	7	8	8
ResNet50	18	18	50	25	17
DenseNet121	30	31	36	56	30
MobileNet	11	14	14	14	61

б/*b*

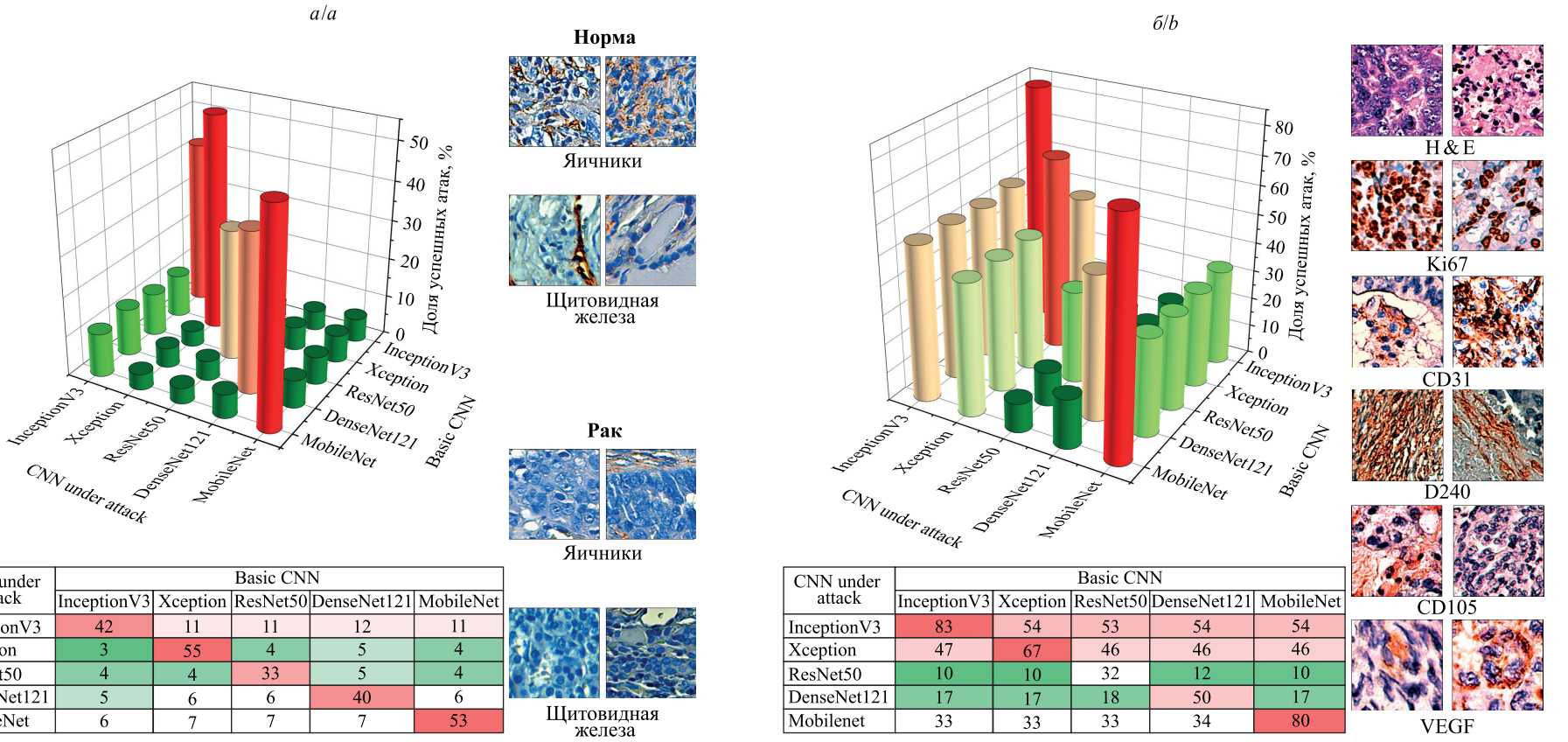




CNN under attack	Basic CNN				
	InceptionV3	Xception	ResNet50	DenseNet121	MobileNet
InceptionV3	25	3	3	2	1
Xception	4	25	2	2	1
ResNet50	2	2	19	3	1
DenseNet121	2	2	2	15	1
MobileNet	2	2	1	2	27

Рис. 4. Доли успешных атак по методу черного ящика при распознавании трехклассового набора рентгеновских изображений грудной клетки X-NR3 (*а*) и двухклассового набора компьютерно-томографических изображений легких CT (*б*)

Fig. 4. The percentage of successful black box attacks for X-NR3 image dataset of X-ray chest images consisting of three classes (*a*) and binary classification of computed tomography image dataset CT of tuberculosis and norm (*b*)



CNN under attack	Basic CNN				
	InceptionV3	Xception	ResNet50	DenseNet121	MobileNet
InceptionV3	42	11	11	12	11
Xception	3	55	4	5	4
ResNet50	4	4	33	5	4
DenseNet121	5	6	6	40	6
MobileNet	6	7	7	7	53

CNN under attack	Basic CNN				
	InceptionV3	Xception	ResNet50	DenseNet121	MobileNet
InceptionV3	83	54	53	54	54
Xception	47	67	46	46	46
ResNet50	10	10	32	12	10
DenseNet121	17	17	18	50	17
Mobilenet	33	33	33	34	80

Рис. 5. Доли успешных атак по методу черного ящика при распознавании набора H-OV-TH гистологических изображений норма/рак по результатам биопсии яичников и щитовидной железы (*а*) и шести классов гистологических изображений, окрашенных различными иммуногистохимическими препаратами (*б*)

Fig. 5. The percentage of successful black box attacks for image dataset H-OV-TH of histology images of the norm and malignant tumors in thyroid glands and the ovary (*a*) and classification images of the H-ST dataset containing histology images stained with the help of six different immuno-histochemical antibodies (*b*)

Для проведения атак по данной методике в настоящей работе рассматриваются пять архитектур глубоких нейронных сетей – InceptionV3, DenseNet121, ResNet50, MobileNet и Xception. Из построенных ранее задач классификации были выбраны две задачи распознавания (классификации) радиологических изображений и две – гистологических. С учетом представленной методики атак по методу черного ящика для каждой выбранной задачи выполнялась следующая последовательность действий.

1. Обучаем каждую из приведенных пяти архитектур сетей.

2. Каждую из обученных сетей атакуем по методу белого ящика, генерируя соответствующее атакующее изображение для каждого исходного изображения из тестовой выборки. Сгенерированные изображения сохраняем.

3. В каждой паре обученных сетей одну сеть назначаем целевой, другую – инструментальной. Проводим атаку на целевую сеть путем подачи атакующих изображений, сгенерированных для инструментальной. Результаты предсказаний классов, выполненных целевой сетью, сохраняем. Затем меняем целевую и инструментальную сети местами и проводим аналогичную атаку.

В качестве алгоритма генерации атакующих изображений был выбран алгоритм PGD с ε, равным 0,1.

Результаты проведения атак. После выполнения описанной последовательности действий для каждой выбранной задачи классификации получаем 25 наборов вероятностей классов изображений (результатов предсказаний), включая 20 наборов для каждой пары *целевая (атакуемая) сеть – инструментальная сеть* и 5 наборов для пар, сети в которых совпадают. Следует отметить, что в последнем случае мы имеем атаки по методу белого ящика, поскольку атаковалась та же самая сеть, для которой генерировались атакующие изображения. По полученным таким образом данным вычислялась доля успешных атак.

На рис. 4 и 5 графически проиллюстрированы результаты проведения экспериментов с атаками на радиологические и гистологические изображения соответственно. Можно видеть, что для двух задач классификации – CT и H-OV-TH – доля успешных атак по методу черного ящика незначительна. Для задачи X-NR3 заметно влияние таких атак, а для задачи H-ST их сила сравнима с силой атак по методу белого ящика (хотя, разумеется, все же меньше). Также стоит отметить, что доля успешных атак по методу черного ящика определяется тем, на вход сети с какой конкретно архитектурой подавались сгенерированные изображения. При этом зависимости от того, для какой сети изображения генерировались, не наблюдается.

Выводы

В данной работе было проведено экспериментальное исследование состязательных атак на глубокие нейронные сети при решении задач классификации биомедицинских изображений различных типов в обоих режимах доступности информации: в режиме белого ящика и в режиме черного ящика. По результатам исследования можно сделать следующие выводы.

1. Проблема состязательных атак актуальна для задач распознавания биомедицинских изображений, поскольку протестированные алгоритмы успешно атакуют обученные нейронные сети так, что их точность падает ниже 15 %.

2. Алгоритм спроецированного градиентного спуска (PGD) при тех же величинах злонамеренных возмущений изображения менее эффективен, чем алгоритм «глубокого обмана» (DeepFool) и алгоритм Карлини – Вагнера (CW).

3. При использовании в качестве метрики сравнения изображений L_2-нормы алгоритмы DeepFool и CW генерируют атакующие изображения близкого качества.

4. В трех из четырех задач распознавания радиологических и гистологических изображений атаки по методу черного ящика с использованием алгоритма PGD показали низкую эффективность.

Библиографические ссылки / References

1. Recht B, Roelofs R, Schmidt L, Shankar V. Do CIFAR-10 classifiers generalize to CIFAR-10? arXiv:1806.00451 [Preprint]. 2018 [cited 2020 August 27]: [25 p.]. Available from: https://arxiv.org/abs/1806.00451.

2. Akhtar N, Mian AS. Threat of adversarial attacks on deep learning in computer vision: a survey. *IEEE Access.* 2018;6:14410–14430. DOI: 10.1109/ACCESS.2018.2807385.

3. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. *Medical Image Analysis.* 2017;42:60–88. DOI: 10.1016/j.media.2017.07.005.

4. Ker J, Wang L, Rao J, Lim T. Deep learning applications in medical image analysis. *IEEE Access.* 2018;6:9375–9389. DOI: 10.1109/ACCESS.2017.2788044.

5. Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. arXiv: 1706.06083v4 [Preprint]. 2017 [cited 2020 August 27]: [28 p.]. Available from: https://arxiv.org/abs/1706.06083.

6. Ozdag M. Adversarial attacks and defenses against deep neural networks: a survey. *Procedia Computer Science.* 2018;140: 152–161. DOI: 10.1016/j.procs.2018.10.315.

7. Wang H, Yu C-N. A direct approach to robust deep learning using adversarial networks. arXiv:1905.09591v1 [Preprint]. 2019 [cited 2020 August 27]: [15 p.]. Available from: https://arxiv.org/abs/1905.09591.

8. Xu W, Evans D, Qi Y. Feature squeezing: detecting adversarial examples in deep neural networks. arXiv:1704.01155v2 [Preprint]. 2017 [cited 2020 August 27]: [15 p.]. Available from: https://arxiv.org/abs/1704.01155.

9. Moosavi-Dezfooli S-M, Fawzi A, Frossard P. DeepFool: a simple and accurate method to fool deep neural networks. arXiv: 1511.04599v3 [Preprint]. 2015 [cited 2020 August 27]: [9 p.]. Available from: https://arxiv.org/abs/1511.04599.

10. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, et al. Intriguing properties of neural networks. In: *2nd International conference on learning representations; 2014 April 14–16; Banff, Canada.* Banff: Springer; 2014. p. 1–10.

11. Carlini N, Wagner D. Towards evaluating the robustness of neural networks. In: *2017 IEEE symposium on security and privacy; 2017 June 26; San Jose, CA, USA.* [S. l.]: IEEE; 2017. p. 39–57. DOI: 10.1109/SP.2017.49.

12. Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. arXiv:1412.6572v3 [Preprint]. 2015 [cited 2020 August 27]: [11 p.]. Available from: https://arxiv.org/abs/1412.6572v3.

Статья поступила в редколлегию 02.09.2020.
Received by editorial board 02.09.2020.