ISSN 1816-0301 (Print)
ISSN 2617-6963 (Online)
УДК 004.9

Поступила в редакцию 10.06.2019
Received 10.06.2019

Принята к публикации 27.06.2019
Accepted 27.06.2019

Экспериментальная оценка состязательных атак на глубокие нейронные сети при решении задач распознавания медицинских изображений

Д. М. Войнов[1], В. А. Ковалев[2]✉

[1]Белорусский государственный университет, Минск, Беларусь
[2]Объединенный институт проблем информатики Национальной академии наук Беларуси, Минск, Беларусь
✉E-mail: vassili.kovalev@gmail.com

Аннотация. Исследуются обнаруженные несколько лет назад проблемы уязвимости глубоких нейронных сетей к так называемым состязательным атакам, которые заставляют сеть принимать ошибочные классификационные решения. Состязательные атаки осуществляются с помощью «атакующих» изображений – незначительно модифицированных версий исходных. Целью работы является изучение зависимости успеха состязательных атак от типа распознаваемых биомедицинских изображений и значений управляющих параметров алгоритмов генерации их атакующих версий. Экспериментальные исследования проводились на примере решения восьми типичных задач медицинской диагностики с использованием глубокой нейронной сети InceptionV3, а также 13 наборов, содержащих более чем 900 000 рентгеновских изображений грудной клетки и гистологических изображений злокачественных опухолей. С увеличением амплитуды вредоносного возмущения и количества итераций генерации зловредного шума вероятность ошибки классификации растет. В то же время различные типы изображений демонстрируют разную чувствительность к данном параметрам. Изображения, которые изначально классифицировались сетью с уверенностью более 95 %, гораздо более устойчивы к атакам. Нейронные сети, обученные для классификации гистологических изображений, оказались более устойчивы к состязательным атакам нежели сети, обученные для классификации рентгеновских изображений.

Ключевые слова: состязательные атаки, глубокое обучение, безопасность нейронных сетей, рентгеновские изображения, гистологические изображения

Для цитирования. Войнов, Д. М. Экспериментальная оценка состязательных атак на глубокие нейронные сети при решении задач распознавания медицинских изображений / Д. М. Войнов, В. А. Ковалев // Информатика. – 2019. – Т. 16, № 3. – С. 14–22.

Experimental assessment of adversarial attacks to the deep neural networks in medical image recognition

Dmitry M. Voynov[1], Vassili A. Kovalev[2]✉

[1]Belarusian State University, Minsk, Belarus
[2]The United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Minsk, Belarus
✉E-mail: vassili.kovalev@gmail.com

Abstract. This paper addresses the problem of dependence of the success rate of adversarial attacks to the deep neural networks on the biomedical image type and control parameters of generation of adversarial examples. With this work we are going to contribute towards accumulation of experimental results on adversarial attacks for the community dealing with biomedical images. The white-box Projected Gradient Descent attacks were

examined based on 8 classification tasks and 13 image datasets containing more than 900 000 chest X-ray and histology images of malignant tumors. An increase of the amplitude and the number of iterations of adversarial perturbations in generating malicious adversarial images leads to a growth of the fraction of successful attacks for the majority of image types examined in this study. Histology images tend to be less sensitive to the growth of amplitude of adversarial perturbations. It was found that the success of attacks was dropping dramatically when the original confidence of predicting image class exceeded 0,95.

Keywords: adversarial attacks, deep learning, security of neural networks, chest X-ray images, histology images

For citation. Voynov D. M., Kovalev V. A. Experimental assessment of adversarial attacks to the deep neural networks in medical image recognition. *Informatics*, 2019, vol. 16, no. 3, pp. 14–22 (in Russian).

Введение. В настоящее время технология глубокого обучения демонстрирует значительные успехи при решении широкого спектра задач обработки, анализа и распознавания изображений. Существенная доля таких задач связана с автоматизацией процессов медицинской диагностики [1, 2], которые включают классификацию изображений для подготовки «второго мнения» в автоматизированных системах компьютеризированной диагностики заболеваний; детектирование подозрительных случаев, требующих участия специалиста при проведении широкомасштабного скрининга населения; генерацию реалистичных искусственных изображений; решение таких вспомогательных задач, как выделение (сегментация) целевых участков медицинских изображений, улучшение их качества и др.

К сожалению, на данный момент основная доля исследований и разработок в области глубокого обучения направлена на достижение глубокими сверточными нейронными сетями (Deep Convolutional Neural Networks, DCNN) максимально возможной точности классификации и распознавания изображений [3]. При этом недостаточно внимания уделяется проблеме безопасности и устойчивости функционирования DCNN-моделей. В результате погоня за единицами процентов прироста точности приводит к тому, что некоторые серьезные проблемы безопасности не получили должного развития как в области исследований, так и в сфере практических разработок.

Настоящая работа посвящена вопросам исследования недавно обнаруженной уязвимости глубоких нейронных сетей [4] к так называемым состязательным атакам (Adversarial Attacks). Проблема уязвимости DCNN является многогранной и представляет значительный научный и практический интерес. Это обусловлено несколькими причинами. Во-первых, модели глубокого обучения неустойчивы к специальным (потенциально злонамеренным) видам модификации входных изображений [5], что весьма небезопасно для систем с повышенной ответственностью. Во-вторых, появилось некоторое экспериментально подтвержденное предположение, что атаки на систему распознавания с помощью специально модифицированных атакующих изображений (Adversarial Examples) могут быть произведены даже при условии, что априорные сведения как об архитектуре, так и об обучающей выборке изображений, использованной при обучении атакуемой DCNN, недоступны.

Поскольку на сегодняшний день информация относительно условий проведения успешных атак ограничена, существует большая потребность в накоплении релевантной информации, которая впоследствии может быть использована при разработке необходимых средств защиты. По очевидным причинам указанная ситуация является особенно острой в области анализа биомедицинских изображений, используемых в процессе диагностики и лечения заболеваний различного характера.

Таким образом, главной целью данного исследования является изучение зависимости успеха состязательных атак на глубокие нейронные сети от таких факторов, как тип анализируемых биомедицинских изображений и значений управляющих параметров алгоритмов генерации их атакующих версий. Для достижения указанных целей необходимо найти ответы на следующие вопросы:

Насколько сильно требуется изменить входное изображение, чтобы глубокая сверточная нейронная сеть допустила ошибку?

При каких условиях нейронная сеть ошибается реже, а при каких чаще?

Зависит ли успешность атаки от типа медицинского изображения и если да, то как?

Характеристика состязательных атак. Состязательной атакой называется некоторое действие, заставляющее нейросетевой классификатор совершать ошибки на этапе применения обученной глубокой нейронной сети. Атаки производятся при помощи атакующих изображений, представляющих собой обычные входные изображения, подлежащие распознаванию, которые были модифицированы специальным образом. Указанные модификации производятся так, что модифицированные изображения визуально практически неотличимы от исходных реальных изображений. Несмотря на это, нейросетевой классификатор расценивает модифицированные изображения как совершенно другие и, как следствие, допускает неприемлемую ошибку. Зачастую алгоритмы генерации атакующих изображений настолько хороши, что различия между атакующим и исходным изображениями незаметны для человеческого глаза. Тем не менее результаты отнесения к соответствующему классу (например, норме или патологии) являются полностью ошибочными.

Собственно, процесс атаки на нейронную сеть заключается в подаче на ее вход сгенерированного атакующего изображения, которое приведет к ошибке распознавания. На сегодняшний день такие атаки в основном осуществляются разработчиками соответствующих программных комплексов в исследовательских целях. Между тем при получении несанкционированного доступа к системе, функционирующей на основе DCNN, появляется возможность проводить атаки на этапе функционирования законченных систем [6], что, разумеется, является серьезной брешью в их безопасности. Несмотря на то что «хорошо атакуются» именно глубокие сверточные сети, состязательные атаки можно проводить и на неглубокие сети с малым количеством слоев. Однако в этом случае эффект будет не таким сильным и, для того чтобы заставить сеть ошибиться, степень отличия атакующих изображений от оригинальных должна быть заметно бóльшей.

Формальное определение атакующего изображения. Пусть $x \in R^D$ – изображение из оригинального набора данных, где $D = n \times m \times c$ – размерность изображения, c – количество каналов. Пусть $F: R^D \to \{1, \ldots, p\}$ – функция, принимающая на вход изображение и возвращающая предсказанный нейронной сетью класс, а p – количество предсказываемых классов. Тогда состязательным примером называется $x^* \in R^D$, такое, что выполняется неравенство

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

и в то же время для некоторого небольшого ϵ соблюдается ограничение

$$\|x^* - x\| \leq \epsilon. \tag{2}$$

Формула (1) показывает, что предсказанный класс изображения x^* отличается от предсказанного класса изображения x. Это достаточно обычное явление, если x^* представляет собой просто некоторое изображение из оригинального набора данных, принадлежащее другому классу. Поэтому для того, чтобы показать суть атакующих изображений, а именно их минимальное отличие от оригинальных изображений, вводится ограничение из формулы (2). Разумеется, при разных значениях ϵ атакующее изображение будет в разной степени отличаться от исходного изображения, в связи с чем выбор рассматриваемой величины ϵ остается за исследователем.

Методы генерации атакующих изображений. Несмотря на то что состязательные атаки впервые были обнаружены относительно недавно, на сегодняшний день существует достаточно большое количество их типов, алгоритмов генерации атакующих изображений и других особенностей, которые описаны ниже.

Классификация атак. В зависимости от доступности информации, необходимой для проведения атаки, алгоритмы генерации делятся на атаки по методу белого и черного ящика. Для атаки по методу белого ящика алгоритму требуется детальная информация об атакуемой нейронной сети, включая ее архитектуру и веса фильтров, полученные в результате обучения. В некоторых случаях необходимы также данные о процедуре тренировки. Для генерации атакующих изображений требуются изображения, принадлежащие к предметной области натренированной сети, например некоторые изображения из обучающей и (или) тестовой выборки.

В случае если исходного набора данных нет, его можно заменить любыми изображениями соответствующей предметной области. Для атаки по методу черного ящика алгоритму требуется нейронная сеть, но уже исключительно как некоторая функция, принимающая на вход изображение и выдающая на выходе вероятности принадлежности его соответствующим классам. Таким образом, в данном случае достаточно иметь только доступ к входным изображениям и вероятностям предсказания на выходе.

По виду ожидаемой ошибки распознавания алгоритмы генерации атак делятся на направленные и ненаправленные. В случае направленных атак атакующие изображения генерируются таким образом, чтобы нейросетевой классификатор ошибочно относил их к некоторому заранее заданному классу, отличному от правильного. В случае ненаправленных атак важно лишь то, чтобы классификатор просто ошибся и отнес изображение к любому из рассматриваемых классов, кроме правильного. Следует отметить, что большинство алгоритмов генерации атакующих изображений имеют версии для атак обоих указанных видов.

Базовая схема алгоритмов генерации атакующих изображений. Несмотря на то что существует достаточно много базовых алгоритмов и эвристик, основные этапы генерации атакующих изображений являются общими и включают три шага:

1. Получение изображения из предметной области атакуемой нейронной сети.
2. Генерирование вредоносного возмущения пикселов исходного изображения.
3. Сложение сгенерированного возмущения со значениями пикселов исходного изображения. Полученный результат и будет являться атакующим изображением (рис. 1).

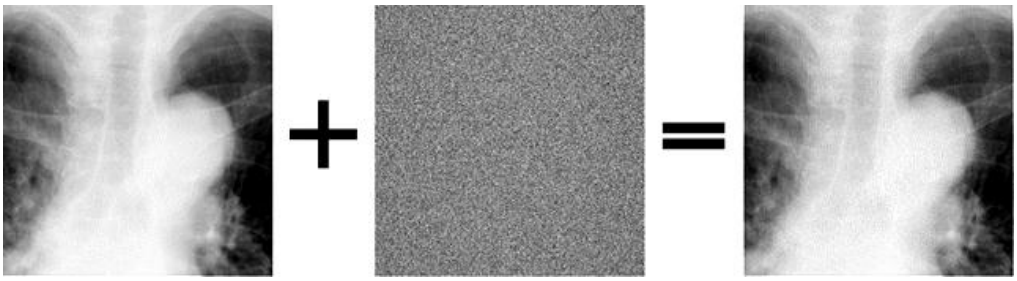


Рис. 1. Общая схема работы алгоритмов генерации атакующих изображений

Легко заметить, что конкретные алгоритмы могут отличаться друг от друга только содержанием первых двух шагов. На шаге 1 различия могут заключаться в источнике оригинального изображения. Им может быть тренировочная выборка, тестовая выборка либо произвольное изображение, предсказываемое сетью. В случае медицинских изображений доступ к изображениям, используемым другими группами разработчиков, как правило, ограничен. На шаге 2 алгоритмы могут отличаться способом генерации вредоносного возмущения. Формальное описание описанной схемы представлено ниже.

Пусть x и x^* – исходное изображение и его атакующая версия, а Δ – вредоносное возмущение. Тогда общая схема атаки выглядит следующим образом:

$$x^* = x + \Delta. \tag{3}$$

Генерация вредоносного возмущения может выполняться с использованием градиентов нейронной сети, рассматриваемой в качестве некоторой функции.

Пусть $y = y(x) = \left(y_1(x), y_2(x), \ldots, y_m(x)\right)$ есть выход нейронной сети как функция входного изображения. Заметим, что все преобразования в современных глубоких нейронных сетях являются дифференцируемыми функциями следующих типов: свертка, нормализация пакета, пулинг, функция активации, полносвязный слой. Исключение составляет разве что популярная активационная функция ReLU, которая не имеет производной в нуле, однако для простоты ее можно доопределить нулем либо единицей. В таком случае функция $y(x)$ становится дифференцируемой, поскольку она является композицией дифференцируемых функций. Как следствие, если m – класс входного изображения, то функция вероятности $y_m(x)$ также дифференцируема. Это послужило мотивацией для использования широко распространенных алгоритмов

локальной оптимизации и в конце концов выродилось в семейство градиентных алгоритмов генерации атакующих изображений. Таким образом, уточняя формулу (1) для градиентных алгоритмов, можно записать

$$\Delta = G\big(x, y(x)\big), \tag{4}$$

где G – некоторая функция, за которой скрывается конкретный алгоритм локальной оптимизации. Стоит отметить, что глобальные оптимизационные алгоритмы также применяются при генерации атакующих изображений (см., например, [7]). Основные градиентные алгоритмы генерации атакующих изображений рассматриваются ниже.

Метод быстрого градиента [7] является одним из первых методов генерации атакующих изображений. Генерация направленного на класс t атакующего изображения зависит от параметра $\alpha > 0$ и определяется следующим образом:

$$x^* = x + \alpha \cdot \operatorname{sign}\big(\nabla y_t(x)\big). \tag{5}$$

Генерация ненаправленного состязательного примера зависит от исходного класса изображения m и определяется как

$$x^* = x - \alpha \cdot \operatorname{sign}\big(\nabla y_m(x)\big). \tag{6}$$

Интуитивно понятен смысл последних двух формул: в случае направленной атаки, двигаясь по знаку градиента, можно ожидать увеличения значения функции $y_t(x)$, т. е. вероятности целевого класса; в случае ненаправленной атаки, двигаясь против знака градиента, можно ожидать уменьшения значения функции $y_m(x)$, т. е. вероятности исходного класса.

Как показала практика, одиночное применение описанного выше возмущения не дает хороших результатов. При малых α атаки редко бывают успешными, при больших – атакующие изображения значительно отличаются от исходных. Поэтому появилась идея использования итеративных методов.

Базовый итеративный метод [7] (Basic Iterative Method, BIM) является усовершенствованной версией метода быстрого градиента посредством добавления механизма итеративности и прямого ограничения применяемого возмущения. В этом случае вместо одиночного вычисления возмущения Δ производится последовательность вычисления возмущений Δ_k при контролируемой величине возмущения.

Генерация направленного на класс t атакующего изображения в базовом и теративном методе зависит от параметров $\alpha > 0$, $n \in N$ и ϵ и определяется следующим образом:

$$x_{k+1} = \operatorname{clip}_{x,\epsilon}\Big(x_k + \alpha \cdot \operatorname{sign}\big(\nabla y_t(x_k)\big)\Big) \tag{7}$$

для $x_0 = x$, $k \in [0, n-1]$ и $x^* = x_n$. Генерация ненаправленного состязательного примера итеративного метода зависит от исходного класса m и определяется как

$$x_{k+1} = \operatorname{clip}_{x,\epsilon}\Big(x_k - \alpha \cdot \operatorname{sign}\big(\nabla y_m(x_k)\big)\Big). \tag{8}$$

В зависимости от механизма определения количества итераций различают два следующих метода: BIM-a, который выполняет фиксированное количество итераций, и BIM-b, который выполняет итерации до тех пор, пока сеть не совершит ошибку при определенном ограничении на количество итераций. Известно, что базовый итеративный метод показывает лучшие результаты по сравнению с предыдущими методами при тех же размерах возмущения. Однако стоит заметить, что каждая итерация этого метода по вычислительной сложности равна одному применению метода быстрого градиента. Следовательно, итеративный метод в целом вычислительно сложнее во столько раз, сколько итераций он применяет.

Следует отметить, что существуют и другие методы генерации атакующих изображений, которые не рассматриваются по причине ограниченности объема настоящей работы.

Известные свойства атакующих изображений. Как уже упоминалось ранее, явление состязательных атак еще не изучено достаточно хорошо. Однако исследователями обнаружены

некоторые воспроизводимые свойства атакующих изображений [4, 8–10]. Ниже перечислены самые, по мнению авторов, значимые из них:

1. Современные глубокие нейронные сети имеют высокую точность в задачах классификации изображений с количеством классов, достигающим 10 000. Поэтому ожидалось, что такие сети будут устойчивы к малым изменениям изображений, но, как показала практика, специальные небольшие, а иногда и неразличимые возмущения входного изображения приводят к тому, что сеть полностью меняет результат предсказания.

2. Атакующие изображения, которые были сгенерированы для проведения атаки на одну модель глубокой сети, зачастую оказываются успешными при атаках на сети других архитектур, обученные на том же наборе изображений или даже на другом наборе из той же предметной области.

3. Несмотря на то что атакующие изображения незначительно отличаются от оригинальных, разница выходов сверточной части нейронной сети (векторов признаков) достаточно большая и может быть даже больше, чем отличия между двумя исходными изображениями, принадлежащими одному, а иногда и разным классам.

Первое приведенное свойство было обнаружено на начальном этапе исследований и, по сути, является «визитной карточкой» состязательных атак. Учитывая минимальную разницу между атакующим и оригинальным изображениями, в сообществе возникли подозрения, что атакующие изображения могут быть не только результатом работы специальных алгоритмов, но и произвольно встречаться в обычных условиях.

Второе свойство используется для проведения состязательных атак по методу черного ящика (о них будет сказано далее), которые являются потенциальными претендентами на роль инструмента хакера моделей глубокого обучения.

Третье свойство показывает, что распространенный в машинном зрении подход к использованию сверточной нейронной сети для выделения признаков также неустойчив и небезопасен. Кроме того, это свойство позволяет предположить, что такая разница векторов признаков появляется не на последнем слое, а постепенно накапливается по мере прохождения данных по сети. Если так, то сделанное предположение может служить возможным объяснением причины успеха метода защиты с использованием JumpReLU активационной функции [11], поскольку она направлена на устранение минимальных отклонений от нуля на выходе пулинг-слоев.

Задачи классификации и используемые исходные изображения. Настоящее исследование проводилось с использованием более 900 000 медицинских изображений двух различных типов: рентгеновских изображений грудной клетки и гистологических изображений образцов мягких тканей, полученных с помощью оптической микроскопии высокого разрешения. Изображения обоих типов широко используются в медицинской практике. В частности, цифровые рентгеновские изображения грудной клетки во многих странах лежат в основе скрининга населения с целью раннего выявления заболеваний легких, сердечно-сосудистой системы, а также аномалий скелета различных видов, а гистологические изображения являются золотым стандартом в диагностике онкологических заболеваний. Краткое описание исходных изображений представлено в таблице.

Характеристики используемых наборов медицинских изображений

Набор данных	Аббревиатура	Задача классификации	Общее количество	Количество в классах	Точность классификации, %
Гистология (метастазы)	H-MT	Норма / метастазы от рака груди	100 000	50 000 в каждом классе	0,97
Гистология (рак яичников и щитовидной железы)	H-OV	Яичники: норма / опухоль	96 000	48 000 в каждом классе	0,92
	H-TH	Щитовидная железа: норма / опухоль	96 000	48 000 в каждом классе	0,94
	H-OV-TH	Яичники, норма / яичники, опухоль / щитовидная железа, норма / щитовидная железа, опухоль	192 000	48 000 в каждом классе	0,91

Окончание таблицы

Набор данных	Аббревиатура	Задача классификации	Общее количество	Количество в классах	Точность классификации, %
Рентген, норма	X-NR2	Возраст: 20–35/ 50–70 лет	200 000	100 000 в каждом классе	0,98
	X-NR3	Возраст: 17–24/ 25–41 /42–80 лет	550 080	183 360 в каждом классе	0,83
Рентген, аорта	X-AO	Норма / аорта развернута	27 000	16 020 / 10 980	0,78
Рентген, туберкулез	X-TB	Норма / туберкулез	28 000	14 000 в каждом классе	0,82

Обучение нейронных сетей. Для каждой из перечисленных выше задач классификации была обучена своя нейронная сеть. В качестве базовой архитектуры нейронных сетей была выбрана широко используемая архитектура InceptionV3. Предварительно обученные на задачах машинного зрения версии архитектуры и веса полученных фильтров не использовались. Для обучения сетей изображения нормировались в интервал [0, 1]. В качестве оптимизатора был выбран AdamOptimizer с одинаковым для каждой задачи обучающим коэффициентом. Во всех случаях с целью достижения приемлемых для исследования точностей достаточно было 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.

Результаты исследования. В настоящей работе в качестве алгоритма генерации атакующих изображений использовался ненаправленный алгоритм спроецированного градиентного спуска по методу белого ящика. Данный метод известен как один из самых сильных среди семейства градиентных методов.

При проведении экспериментов амплитуда вредоносного градиентного возмущения пикселов варьировалась в пределах 0,02...0,2 от максимального значения пикселов с шагом 0,02 (рис. 2). Примеры успешных атак на гистологические изображения, используемые при диагностике онкологических заболеваний, приведены на рис. 3. Видно, что атакующие версии изображений (справа в каждой паре) визуально являются весьма близкими к реальным и тем не менее все они были ошибочно отнесены обученной глубокой сетью с архитектурой Inception v3 к противоположному классу.

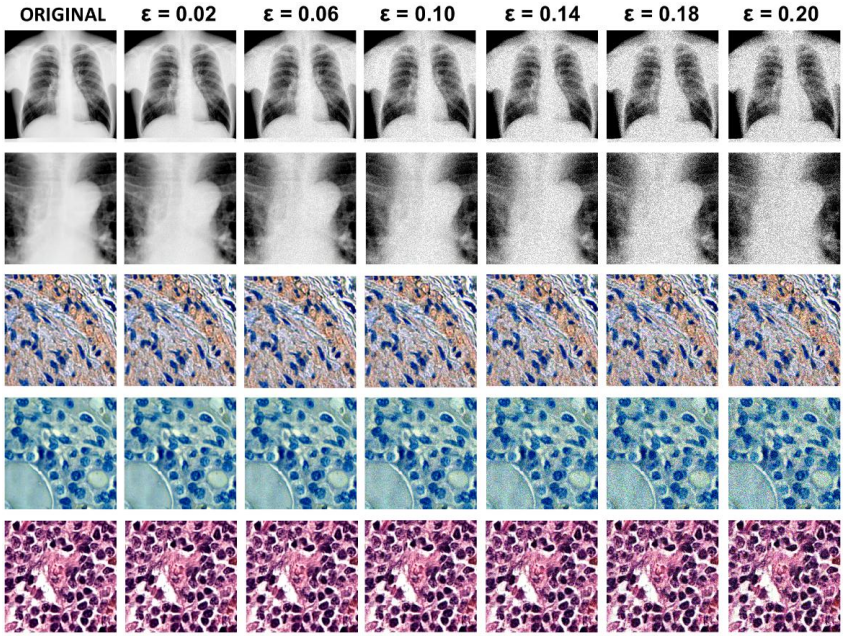
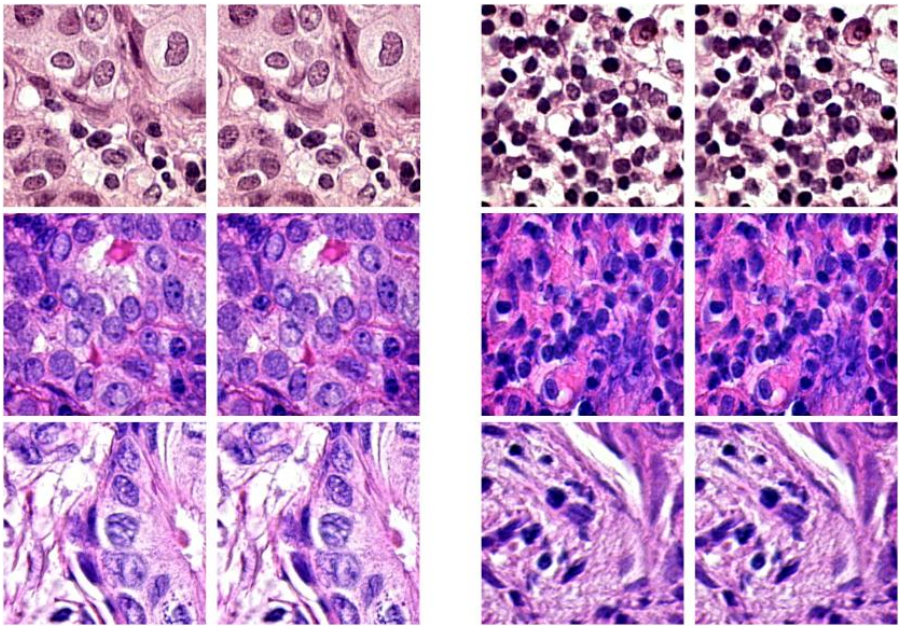


Рис. 2. Примеры исходных рентгеновских (первые две строки) и гистологических (последующие три строки) изображений с различной величиной вредоносного возмущения ε



а) *б)* *в)* *г)*

Рис. 3. Примеры успешных атак на гистологические изображения: исходные изображения участков злокачественных опухолей (*а*) и их атакующие версии (*б*), ошибочно распознанные нейронной сетью как нормальная ткань; исходные изображения нормальной ткани (*в*) и их атакующие версии (*г*), ошибочно распознанные как опухоль

Детальные результаты исследования влияния изучаемых факторов на успешность состязательных атак, которые проводились в рамках исследуемых в данной работе задач классификации медицинских изображений, представлены в виде графиков на рис. 4.

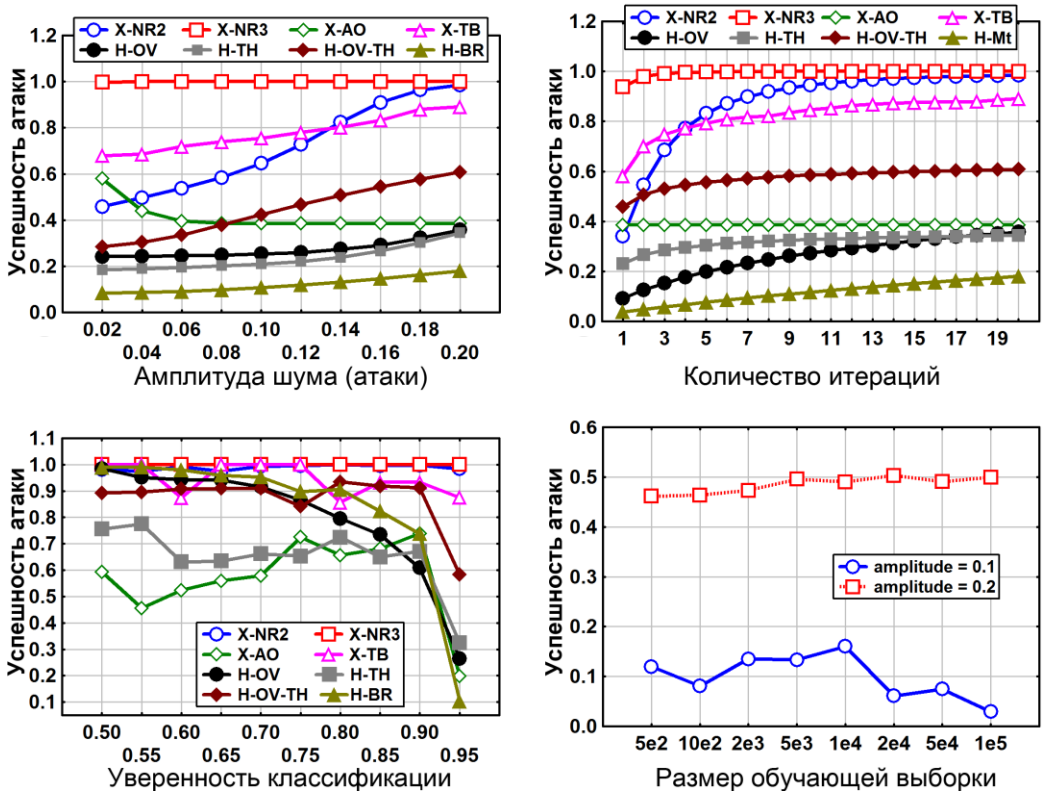


Рис. 4. Результаты оценки влияния различных факторов на успешность атаки

Заключение. Результаты, полученные в настоящей работе, позволяют сделать следующие выводы:

1. С увеличением амплитуды вредоносного возмущения вероятность ошибки предсказания состязательного примера растет. Однако различные типы изображений демонстрируют разную чувствительность к данному параметру (вплоть до противоположной).

2. Рост количества итераций метода спроецированного градиентного спуска постепенно увеличивает количество ошибок, допускаемых сетью, с асимптотической сходимостью к некоторому максимальному значению.

3. Изображения, которые классифицируются сетью с уверенностью более 95 %, гораздо более устойчивы к состязательным атакам.

4. Нейронные сети, обученные для классификации гистологических изображений, оказались более устойчивы к злонамеренным атакам, нежели сети, обученные для классификации рентгеновских изображений грудной клетки.

Авторы полагают, что полученные экспериментальные данные расширяют имеющиеся знания относительно состязательных атак в области анализа медицинских изображений.

References

1. Litjens G., Kooi T., Bejnordi B., Setio A., Ciompi F., Ghafoorian M. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 2017, vol. 42, pp. 60–88.

2. Ker J., Wang L., Rao J., Lim T. Deep learning applications in medical image analysis. *IEEE Access*, 2018, vol. 6, pp. 9375–9389.

3. Recht B., Roelofs R., Schmidt L., Shankar V. *Do CIFAR-10 Classifiers Generalize to CIFAR-10?* ArXiv.org, 2018. Available at: https://arxiv.org/abs/1806.00451 (accessed 15.05.2019).

4. Szegedy C., Wojciech Z., Sutskever I., Bruna J., Dumitru E., Goodfellow I., Fergus R. Intriguing properties of neural networks. *International Conference on Learning Representations (ICLR'2014), Banff, Canada, 14–16 April 2014*. Banff, 2014, pp. 1–10.

5. Akhtar N., Mian A. S. Threat of adversarial attacks on deep learning in computer vision. *IEEE Access*, 2018, vol. 6, pp. 14 410–14 430.

6. Papernot N., McDaniel P., Goodfellow I., Jha S., Berkay Celik Z., Swami A. *Practical Black-Box Attacks agains Machine Learning*. ArXiv.org, 2017. Available at: https://arxiv.org/abs/1602.02697 (accessed 15.05.2019).

7. Xu W., Evans D., Qi Y. *Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks*. ArXiv.org, 2017. Available at: https://arxiv.org/abs/1704.01155 (accessed 15.05.2019).

8. Goodfellow I., Shlens J., Szegedy C. *Explaining and Harnessing Adversarial Examples*. ArXiv.org, 2015. Available at: https://arxiv.org/abs/1412.6572 (accessed 15.05.2019).

9. Madry A., Makelov A., Schmidt L., Tsipras D., Vladu A. *Towards Deep Learning Models Resistant to Adversarial Attacks*. ArXiv.org, 2017. Available at: https://arxiv.org/abs/1706.06083 (accessed 15.05.2019).

10. Ozdag M. Adversarial attacks and defenses against deep neural networks: a survey. *Procedia Computer Science*, 2018, vol. 140, pp. 152–161.

11. Ericson N. B., Yao Z., Mahoney W. *JumpReLU: A Retrofit Defense Strategy for Adversarial Attacks*. ArXiv.org, 2019. Available at: https://arxiv.org/abs/1904.03750 (accessed 15.05.2019).

Информация об авторах

Войнов Дмитрий Михайлович, магистрант, Белорусский государственный университет, Минск, Беларусь.
E-mail: voynovdd@gmail.com

Ковалев Василий Алексеевич, кандидат технических наук, заведующий лабораторией анализа биомедицинских изображений, Объединенный институт проблем информатики Национальной академии наук Беларуси, Минск, Беларусь.
E-mail: vassili.kovalev@gmail.com

Information about the authors

Dmitry M. Voynov, Master Student, Student of Belarusian State University, Minsk, Belarus.
E-mail: voynovdd@gmail.com

Vassili A. Kovalev, Cand. Sci. (Eng.), Head of the Laboratory of Biomedical Images Analysis, The United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Minsk, Belarus.
E-mail: vassili.kovalev@gmail.com