DOI: 10.51790/2712-9942-2021-2-3-4

ОШИБКИ МАШИННОГО ОБУЧЕНИЯ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ И ИСПОЛЬЗОВАНИЕ ИХ ДЛЯ АТАК НА СИСТЕМЫ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Т. В. Гавриленко[1,2], **А. В. Гавриленко**[2]
[1] *Сургутский филиал Федерального государственного учреждения «Федеральный научный центр Научно-исследовательский институт системных исследований Российской академии наук», г. Сургут, Российская Федерация, taras.gavrilenko@gmail.com*
[2] *Сургутский государственный университет, г. Сургут, Российская Федерация, gavrilenko.anna.v@gmail.com*

Аннотация: в статье приведен обзор различных методов атак и подходов к атакам на системы искусственного интеллекта, построенных на основе искусственных нейронных сетей. Показано, что начиная с 2015 года исследователи в различных странах активно развивают методы атак и подходы к атакам на искусственные нейронные сети, при этом разработанные методы и подходы могут иметь критические последствия при эксплуатации систем искусственного интеллекта. Делается вывод о необходимости развития методологической и теоретической базы искусственных нейронных сетей и невозможности создания доверительных систем искусственного интеллекта в текущей парадигме.

Ключевые слова: искусственные нейронные сети, ошибки машинного обучения, атаки на системы искусственного интеллекта.

Благодарности: работа выполнена при поддержке РФФИ, грант № 18-47-860004.

Для цитирования: Гавриленко Т. В., Гавриленко А. В. Ошибки машинного обучения искусственных нейронных сетей и использование их для атак на системы искусственного интеллекта. *Успехи кибернетики*. 2021;2(3):23–32. DOI: 10.51790/2712-9942-2021-2-3-4.

EXPLOITING ARTIFICIAL NEURAL NETWORKS MACHINE LEARNING ERRORS FOR ATTACKS ON AI SYSTEMS

T. V. Gavrilenko[1,2], **A. V. Gavrilenko**[2]
[1] *Surgut Branch of Federal State Institute "Scientific Research Institute for System Analysis of the Russian Academy of Sciences", Surgut, Russian Federation, taras.gavrielnko@gmail.com*
[2] *Surgut State University, Surgut, Russian Federation, gavrilenko.anna.v@gmail.com*

Abstract: the paper provides an overview of methods and approaches to attacks on neural network-based artificial intelligence systems. It is shown that since 2015, global researchers have been intensively developing methods and approaches for attacks on artificial neural networks, while the existing ones may have critical consequences for artificial intelligence systems operations. We come to the conclusion that theory and methodology for artificial neural networks is to be elaborated, since trusted artificial intelligence systems cannot be created in the framework of the current paradigm.

Keywords: artificial neural networks, machine learning errors, attacks on artificial intelligence systems.

Acknowledgements: this study is supported by RFBR grant 18-47-860004.

Cite this article: Gavrilenko T. V., Gavrilenko A. V. Exploiting Artificial Neural Networks Machine Learning Errors for Attacks on AI Systems. *Russian Journal of Cybernetics*. 2021;2(3):23–32. DOI: 10.51790/2712-9942-2021-2-3-4.

В 2015 году группа разработчиков систем искусственного интеллекта компании Google впервые показала применение искусственных нейронных сетей с состязательным обучением для искажения исходного изображения таким образом, чтобы объекты на изображении идентифицировались неправильно. Это одна из самых знаковых работ, в которой показано, как исказить пиксели изображения, чтобы классификатор принял ошибочное решение. В основе метода лежит факт, что изображения часто представлены в виде 8-битных значений (каждый пиксель может иметь одно целочисленное значение в диапазоне от 0 до 255, то есть в сумме 2^8 значений). Следовательно, если искажения не превышают минимального значения, которое может быть представлено в изображении, то классификатор должен полностью их игнорировать и считать искаженное изображение неизмененным. Но авторы показывают, что это не так [2].

В статье [2] авторы определяют ошибочную классификацию входных данных уравнением:

$$w^T \tilde{x} = w^T x + w^T \eta,$$

где $\tilde{x}$ — входные данные, предназначенные для введения нейросети в заблуждение; $w^T x$ — выходные данные классификатора по неизмененному изображению (которое классифицируется корректно); η — это специальный вектор, добавленный к исходным входным данным таким образом, чтобы вся сеть приняла ошибочное решение о классификации.

То есть уравнение читается так: «Сеть может ошибиться в классификации, если к оригинальным входным данным добавлены такие данные, что получившийся результат заставляет нейросеть отнести его к другому классу». Это совершенно очевидно. Куда интереснее, как находить значение η [1], которое определяется следующим образом:

$$\eta = \epsilon \ \mathrm{sign}(\nabla_x J(\theta, x, y)) \ ,$$

где ∇_x — это градиенты входных данных; $J(\theta, x, y)$ — функция стоимости, используется при обучении искусственных нейронных сетей; θ — параметры модели; x — входные данные; y — целевые выходные данные, то есть «ошибочный» класс.

Поскольку вся сеть является дифференцируемой, значения градиента можно легко найти с помощью метода обратного распространения ошибки.

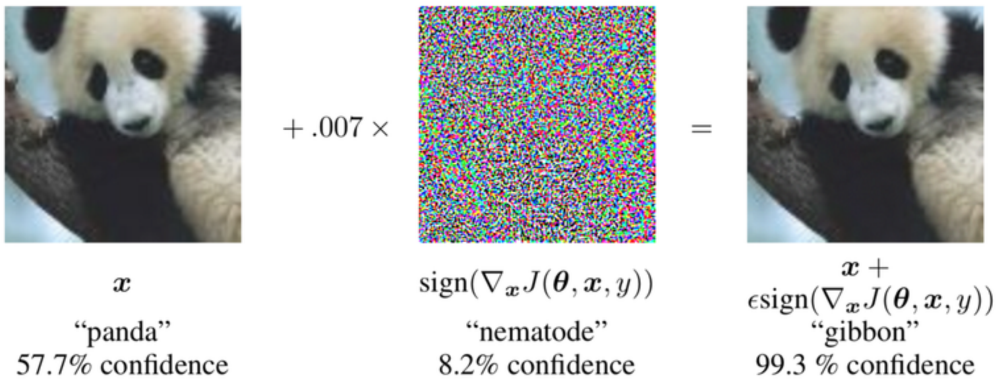

Рис. 1. *Демонстрация быстрой генерации состязательных примеров применительно к GoogLeNet на ImageNet. Добавляется очень маленький вектор, элементы которого равны знаку элементов градиента функции стоимости по отношению к входным данным, классификация изображений в GoogLeNet дает ошибочный результат. Здесь 0,007 соответствует величине самого маленького бита 8-битного кодирования изображения после преобразования GoogLeNet в действительные числа [1]*

Исследователи из Колумбийского и Лихайского университетов в 2017 году разработали систему для тестирования алгоритмов глубокого обучения. С помощью этой системы разработчики проверили искусственные нейронные сети, в том числе управляющие беспилотными автомобилями, и обнаружили в них тысячи ошибок, проявляющихся в необычных для искусственных нейронных сетей условиях [2]. Примеры возникающих ошибок представлены ниже на рис. 2. Наиболее критичными являются два левых столбца, в которых показаны ошибки принятия решения системой беспилотного управления автомобилем.

Разработчики протестировали свой алгоритм под названием DeepXplore с помощью больших наборов данных для тренировки алгоритмов разного назначения и выявили серию ошибок, которые возникают в процессе распознавания образов и принятия решения искусственными нейронными сетями [2]. В работе показаны как ошибки в распознавании отдельных объектов, так и ситуационное принятие решения искусственными нейронными сетями.

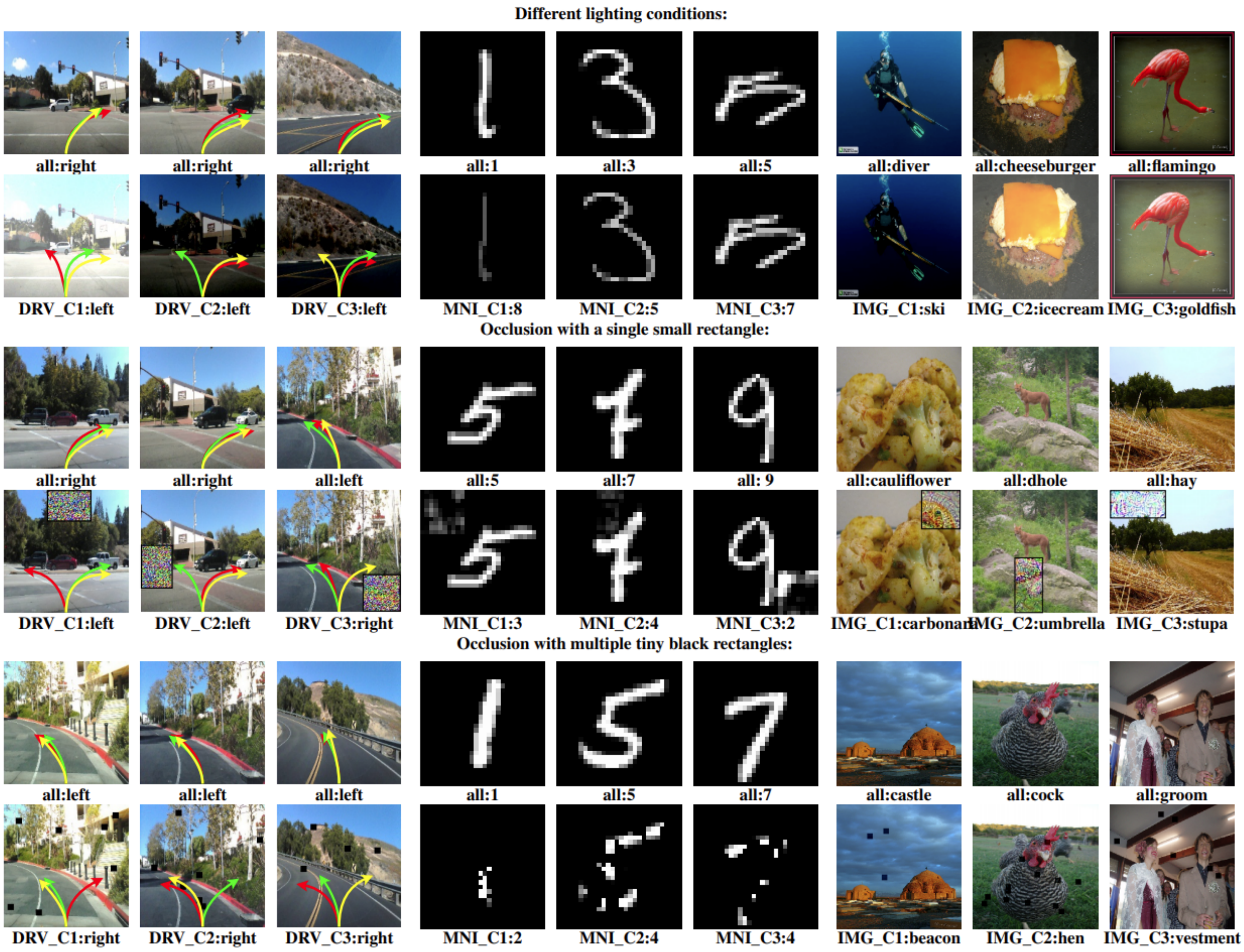

Рис. 2. *Пример разных результатов, которые выдают программы в ответ на одни и те же входные данные с внедренными шумовыми компонентами. Нечетные строки — исходные изображения и результаты распознавания, четные строки — реакция на изменение результатов работы нейронной сети после введения шума [2]*

В 2017 году группа исследователей в статье [3] показала эффективность методики атаки на искусственные нейронные сети при помощи одного единственного пикселя, интегрированного в изображение. В статье представлены не только практические результаты, но и методика атаки на искусственные нейронные сети с использованием одного единственного символа в исходном изображении. На рис. 3 показано, как на основе методики авторов им удается менять опознаваемый объект при помощи всего лишь одной интегрированной точки, кроме этого, данная группа разработчиков показала, что атака может быть целевой, а не случайной.

Механизм атаки и его модификации следующий: создание состязательных изображений можно формализовать как задачу оптимизации с ограничениями. Авторы предполагают, что входное изображение может быть представлено вектором, в котором каждый скалярный элемент — один пиксель. Пусть f будет классификатором целевого изображения, который получает n-мерные входные данные, $x = (x_1, \dots, x_n)$ будет исходным естественным изображением, правильно классифицированным как класс t. Следовательно, вероятность того, что x принадлежит классу t, равна $f_t(x)$. Вектор $e(x) = (e_1, \dots, e_n)$ является аддитивным противодействующим возмущением согласно x, целевому классу adv и ограничению максимальной модификации L. L всегда измеряется длиной вектора $e(x)$. Цель злоумышленников в случае целевых атак — найти оптимизированное решение $e(x)^*$ для следующего выражения:

$$\begin{aligned} &\underset{e(x)^*}{\text{maximize}} && f_{adv}(x + e(x)) \\ &\text{subject to} && \|e(x)\| \le L. \end{aligned}$$

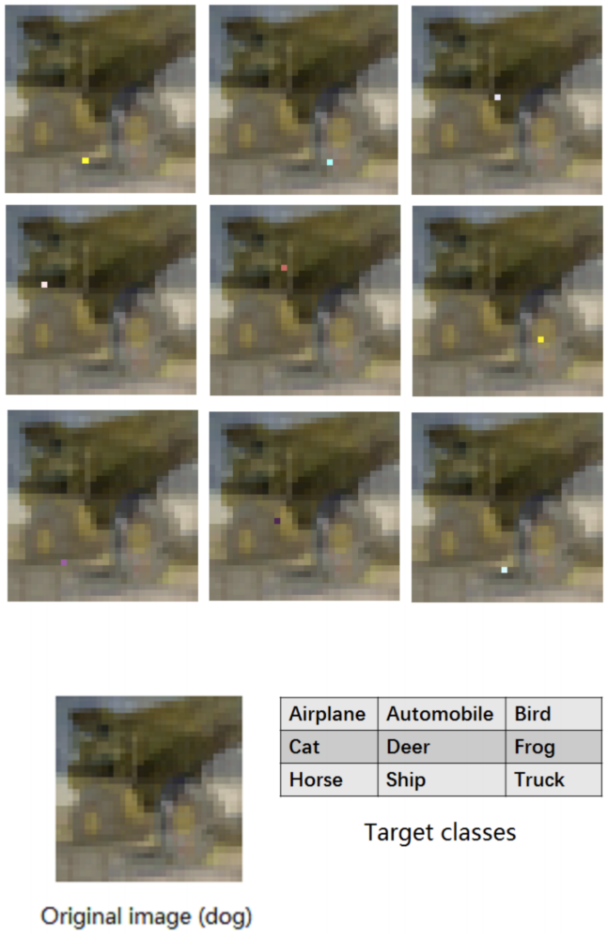

Рис. 3. *Пример атаки на искусственную нейронную сеть при помощи внедрения одной точки. Показано, что при использовании специальной методики можно менять пространство состояний не случайным образом, а в соответствии с требованиями [3]*

Проблема заключается в нахождении двух значений: (а) в каких измерениях необходимо внести возмущение и (б) уровень модификации значения. Авторы предложили свой подход:

$$\begin{aligned} &\underset{e(x)^*}{\text{maximize}} && f_{adv}(x + e(x)) \\ &\text{subject to} && ke(x)k \leq d, \end{aligned}$$

где d — небольшое число.

В случае однопиксельной атаки d = 1. Фактически однопиксельное возмущение позволяет изменять изображение в выбранном направлении из n возможных направлений с произвольной силой [3].

2017 год был показательным с точки зрения появления различных механизмов атак на нейронные сети, при этом одной из самых знаковых была публикация специалистов MIT «Query-efficient Black-box Adversarial Examples» («Примеры состязательных черных ящиков с эффективными запросами») [4]. Специалистами было продемонстрировано, что могут быть взломаны практически любые, даже самые сложные ИИ с использованием алгоритма, основанного на естественных эволюционных стратегиях (NES), который позволяет генерировать состязательные примеры в настройках черного ящика без обучения замещающей сети [4]. При этом было доказано, что предлагаемый исследователями метод на два порядка эффективнее всех предыдущих, быстродействие на некоторых примерах возросло до 1000 раз. Представленные в статье [4] примеры крайне показательны (рис. 4). Верхняя строка иллюстрации показывает результат идентификации объектов на исходной фотографии, ниже представлены результаты.

На рис. 4 можно увидеть, как внедрение целевого возмущения в объеме $\varepsilon = 0.1$ приводит к полной потере способности системы к распознаванию образов, основанной на искусственных нейрон-

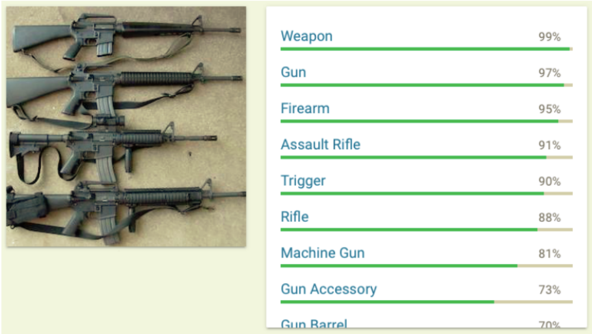

Figure 6. The Google Cloud Vision Demo labelling on the unperturbed image.

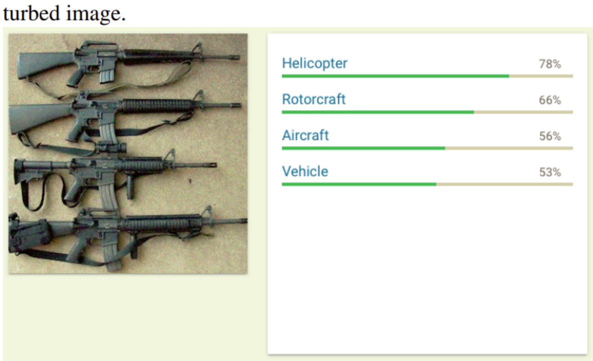

Figure 7. The Google Cloud Vision Demo labelling on the adversarial image generated with ℓ_∞ bounded perturbation with $\epsilon = 0.1$: the original class is no longer a returned label.

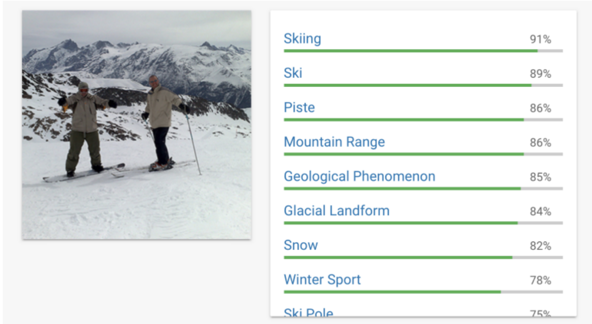

Figure 8. The Google Cloud Vision Demo labelling on the unperturbed image.

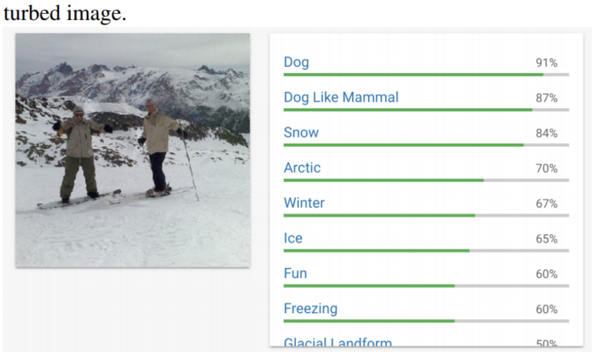

Figure 9. The Google Cloud Vision Demo labelling on the adversarial image generated with ℓ_∞ bounded perturbation with $\epsilon = 0.1$: the image is labeled as the target class.

Рис. 4. *Пример атаки на искусственную нейронную сеть при помощи эволюционных алгоритмов (NES) [4]. Показатель уровня внедренного возмущения в фотографию* $\varepsilon = 0.1$

ных сетях, выполнять свои прямые задачи. Пространство состояний существенно сдвинуто. Причем модифицированные фотографии с точки зрения человека практически равнозначны и не вызывают никаких трудностей в идентификации объектов на них.

Кроме того, в 2017–2018 годах много шума наделали работы, посвященные так называемым вредоносным заплаткам (adversarial patch). Наиболее знаковая и, можно сказать, первая подробная работа, посвященная данной проблеме, была озвучена на 31-й конференции по нейронным системам обработки информации (Conference on Neural Information Processing Systems (NIPS 2017)). Исследователи корпорации Google представили метод создания универсальных, надежных и целевых заплаток (adversarial patch), состязательных изображений для реального мира. Вредоносные заплатки обладают универсальностью, потому что их можно использовать для атаки на любую сцену в окружающем мире. Заплатки работают с широким спектром преобразований и заставляют классификатор выводить любой целевой класс вместо правильного. Эти состязательные заплатки можно распечатать, добавить к любой сцене, сфотографировать и представить классификаторам изображений, даже когда заплатки небольшие. Они заставляют классификаторы игнорировать другие элементы сцены и менять целевой класс [5]. Пример атаки с использованием физической вредоносной заплатки представлен на рис. 5.

В статье был представлен и метод генерации таких заплаток, а также функция оптимизации для создания подобных элементов. Данная атака — одна из наиболее серьезных по причине, что интеграция вредоносного воздействия осуществляется не в цифровое изображение или любой другой массив данных, а в реальном мире. В частности, для того, чтобы обмануть систему распознавания и идентификации объектов, не надо пытаться воздействовать на результат оцифровки данных, можно просто в реальном мире разместить подготовленное изображение, и система уже не сможет идентифицировать и распознать объекты. Можно выйти в футболке с некоторым изображением, и все системы автоматического управления с распознаванием объектов с помощью нейронных сетей будут совершать фатальные ошибки, как только данное изображение будет попадать в поле видимости этих систем. Вредоносные заплатки могут размещаться в любых местах и иметь различное назначение.

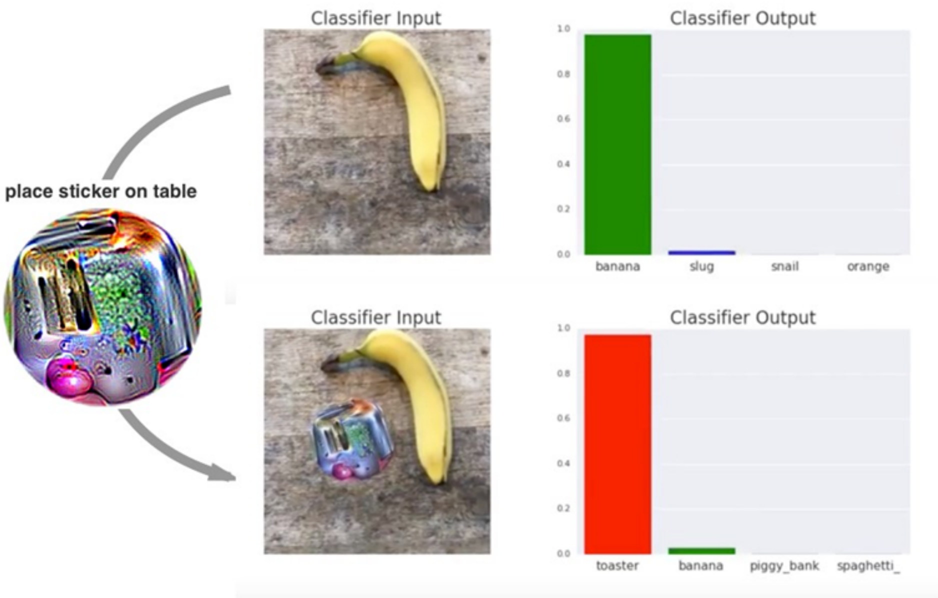

Рис. 5. *Реальная атака на VGG16 с использованием физической вредоносной заплатки. Фотография стола с бананом (верхняя фотография) проходит через VGG16, сеть сообщает о классе «банан» с достоверностью 97 % (верхний график). Если физически разместить на столе наклейку, предназначенную для класса «тостер» (нижняя фотография), фотография будет классифицирована как тостер с достоверностью 99 % (нижний график) [5]*

Крайне интересен еще один вид атаки, который сводит на нет биометрические системы распознавания голоса. В 2017 году исследователи из Калифорнийского университета продемонстрировали возможность применения атаки, описанной разработчиками корпорации Google в 2015 году для систем распознавания голоса. Результаты работы были также озвучены на 31-й конференции по нейронным системам обработки информации (Conference on Neural Information Processing Systems (NIPS 2017)). В краткой статье авторы представили первую в своем роде демонстрацию состязательных атак на модель классификации речи. Алгоритм выполняет целевые атаки с успехом 87 % путем добавления небольшого фонового шума без знания основных параметров модели и архитектуры искусственной нейронной сети распознавания речи. Атака изменяет только наименее значимые биты подмножества элементов цифрового аудиосигнала, а шум не меняет 89 % восприятия цифрового аудиосигнала слушателем-человеком [6]. Ниже представлена иллюстрация авторов исследования (рис. 6).

Рис. 6. *Состязательные атаки на речевые системы распознавания: злоумышленник добавляет к звуку небольшой шум, так, что он ошибочно классифицируется моделью распознавания речи, но не меняет человеческого восприятия [6]*

Данное исследование показывает, что могут быть атакованы системы биометрической идентификации личности и, что еще немаловажно, системы, реализующие голосовое управление. Сейчас можно часто услышать, что нельзя при разговоре с сомнительным абонентом произносить те или иные слова, чтобы случайно не подтвердить некую операцию, например, с банковским счетом, тем более, что сейчас все больше и больше вместо операторов-людей используют голосовые помощники, включающие в себя системы распознавания речи. В результате неважно, кто и что говорит, систему можно заставить слышать то, что нужно злоумышленнику! Мало того, любой внешний наблюдатель-человек будет слышать абсолютно другое!

В 2019 году появились работы, в которых исследователи представляли различные алгоритмы борьбы с искажениями и изображениями, в которые внедрены искажения. Изображения с внедренными искажениями принято называть DeepFake. В данном случае составное слово содержит часть названия наиболее прогрессивного способа обучения искусственных нейронных сетей Deep Learning (глубокое обучение) и Fake (обман). Появились даже инструменты борьбы с DeepFake. Но в 2020 году начали публиковаться работы, в которых авторы описывают механизмы обхода/обмана искусственных нейронных сетей. Фактически разработчики из Калифорнийского университета предложили модификацию метода, представленного еще в 2015 году [1]. Если говорить о методе, который использовали исследователи, то его суть можно сформулировать следующим образом. Если мы имеем изображение, которое однозначно не является правдоподобным, тогда путем повторного наложения шума, сгенерированного специализированной нейронной сетью, можно перевести его из ряда ошибочных в «достоверные». Иллюстрация работы алгоритма от разработчиков Калифорнийского университета представлена на рис. 7.

В 2020 году большой интерес вызвала библиотека DeepRobust для PyTorch (фреймворк машинного обучения для языка Python с открытым исходным кодом, созданный на базе Torch). PyTorch используется для решения различных задач: компьютерное зрение, обработка естественного языка, разрабатывается преимущественно группой искусственного интеллекта Facebook. На текущий момент в DeepRobust доступны следующие методы атак на модели, которые работают с картинками:

1) LBFGS attack;
2) FGSM attack;
3) PGD attack;
4) DeepFool attack;
5) CW attack;
6) One pixel attack;
7) BPDA attack;
8) Universal attack;
9) Nattack.

Среди доступных методов защиты от состязательных атак:

1) FGSM обучение;
2) Fast: улучшенная версия FGSM обучения;
3) PGD обучение;
4) YOPO: улучшенная версия PGD обучения;
5) TRADES;
6) Thermometer Encoding;
7) LID-based классификатор состязательных примеров.

Библиотека поддерживает стандартные наборы данных компьютерного зрения: MNIST, CIFAR-10 и ImageNet. Кроме того, поддерживаемые архитектуры моделей включают в себя SimpleCNN и ResNet (ResNet18, ResNet34, ResNet50).

Также доступны следующие методы атак на модели, которые обрабатывают графы:

1) Nettack;
2) FGA;
3) Mettack;
4) RL-S2V;
5) PGD, Min-max;
6) DICE;
7) IG-Attack;
8) NIPA;
9) RND.

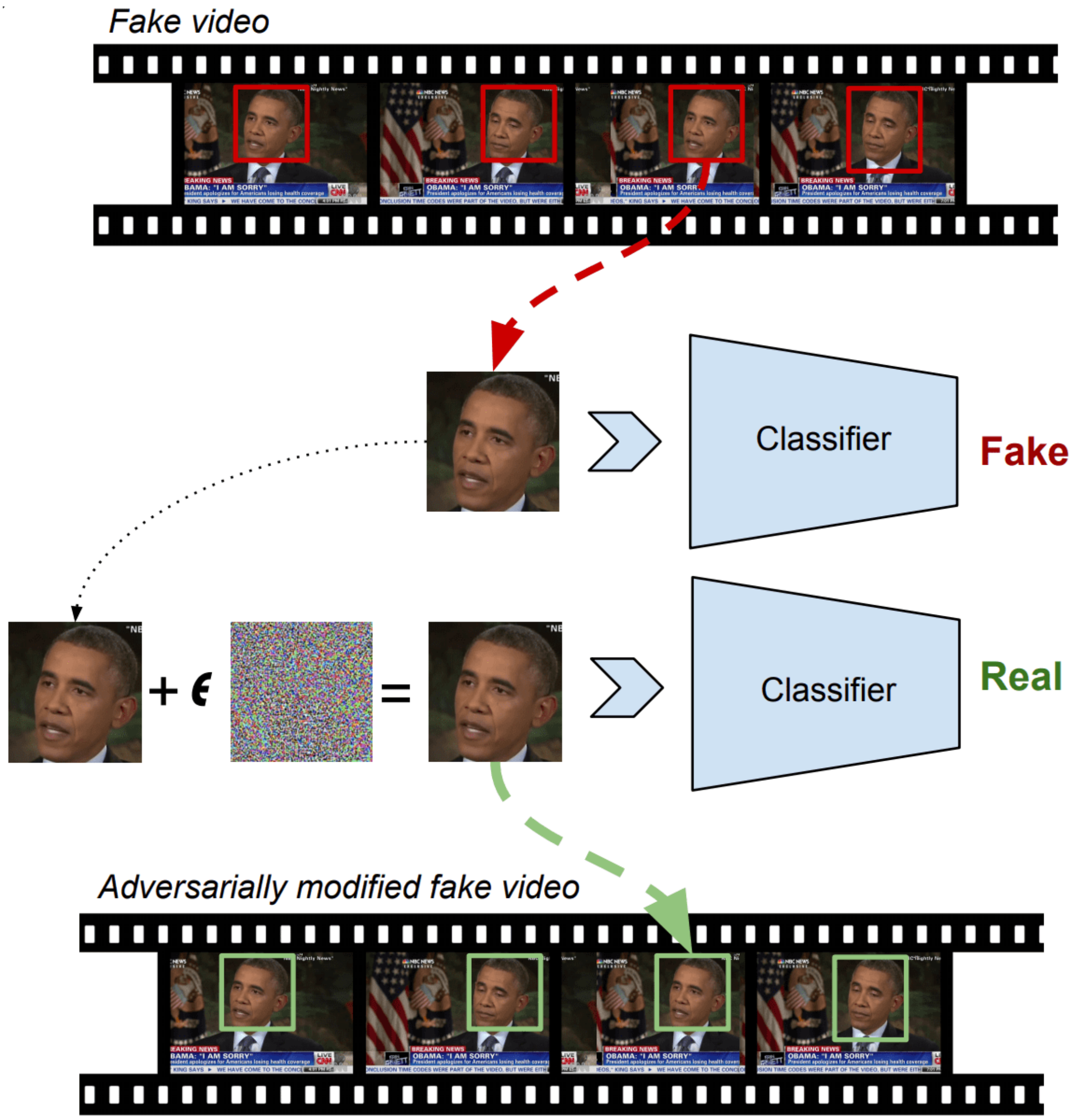

Рис. 7. *Название рисунка*

Среди доступных методов защиты от состязательных атак:

1) RGCN;
2) GCN-Jaccard;
3) GCN-SVD;
4) Adv-training [8].

Последнее время количество публикаций по обману искусственных нейронных сетей резко возросло. Компьютерные инженеры из Чикагского университета разработали инструмент Fawkes, который маскирует фотографии для защиты от систем распознавания лиц, используемые различными ИТ-гигантами для сбора информации.

В завершение обзора хотелось бы привести еще одну знаковую статью, опубликованную на https://arxiv.org/, – «Adversarial Patch Camouflage Against Aerial Detection» («Камуфляж вредоносной заплаткой против обнаружения объектов с воздуха») [9]. В данной работе авторы показали в сравнении методы классического и цифрового камуфляжа (рис. 8).

На рис. 9 показана эффективность цифрового камуфляжа от автоматического распознавания летательных аппаратов.

Выводы

Искусственные нейронные сети обладают огромным потенциалом для решения различных за-

(a) Camouflage net *(b) Patch camouflage*

Рис. 8. *Маскировочная сетка, используемая для сокрытия объекта (слева), и самолет с вредоносной заплаткой (цифровой камуфляж), которая может скрыть его от автоматического обнаружения (справа) [9]*

Рис. 9. *Изображения реактивных истребителей с наложением обнаружений на доверительный порог 0,4. Достоверность обнаружения реактивных истребителей находится в диапазоне от 0,45 до 0,78, при этом достоверность необнаруженного реактивного самолета составляет 0,23. Истребители с вредоносными заплатками не обнаруживаются на этом пороге: доверительный интервал от 0,01 до 0,04 с единственным пиковым значением 0,14 [9]*

дач, но при этом имеют ряд проблемных ограничений в текущей методологической и теоретической базе. Отсутствие четких теоретических представлений и технических разработок, направленных на создание «доверенных систем», приводит к неутешительному выводу о существенных ограничениях использования систем искусственного интеллекта, построенных на базе искусственных нейронных сетей, особенно в критических областях хозяйственной деятельности.

ЛИТЕРАТУРА

1. Goodfellow I. J., Shlens J., Szegedy C. Explaining and Harnessing Adversarial Examples. *Proceedings of the International Conference on Learning Representations (ICLR)*. 2015.
2. Kexin Pei, Yinzhi Cao, Junfeng Yang, Suman Jana. DeepXplore: Automated Whitebox Testing of Deep Learning Systems. *SOSP '17*, October 28, 2017, Shanghai, China.
3. Jiawei Su, Danilo Vasconcellos Vargas, Kouichi Sakurai. *One Pixel Attack for Fooling Deep Neural Networks*. 2017. Режим доступа: https://arxiv.org/pdf/1710.08864.pdf.
4. Andrew Ilyas, Logan Engstrom, Anish Athalye, Jessy Lin. *Query-efficient Black-box Adversarial Examples*. 2017. Режим доступа: https://arxiv.org/pdf/1712.07113.pdf.
5. Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, Justin Gilmer. Adversarial Patch. *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA. 2017. Режим доступа: https://arxiv.org/pdf/1712.07113.pdf.

6. Moustafa Alzantot, Bharathan Balaji, Mani Srivastava. Did You Hear That? Adversarial Examples Against Automatic Speech Recognition. *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA. 2017. Режим доступа: https://arxiv.org/pdf/1801.00554v1.pdf.
7. Shehzeen Hussain, Paarth Neekhara, Malhar Jere, Farinaz Koushanfar, Julian McAuley. Adversarial Deepfakes: Evaluating Vulnerability of Deepfake Detectors to Adversarial Examples. *WACV 2021*. Режим доступа: https://adversarialdeepfakes.github.io/.
8. *DeepRobust: библиотека для состязательных атак на нейросети*. Режим доступа: https://neurohive.io/ru/frameworki/deeprobust-biblioteka-dlya-sostyazatelnyh-atak-na-nejroseti/.
9. Ajaya Adhikari, Richard den Hollander, Ioannis Toliosa, Michael van Bekkuma, Anneloes Balc, Stijn Hendriksc, Maarten Kruithofb, Dennis Grossd, Nils Jansend, Guillermo Péreze, Kit Buurmana, Stephan Raaijmakersa. *Adversarial Patch Camouflage against Aerial Detection*. 2020. Режим доступа: https://arxiv.org/pdf/1712.07113.pdf.