ISSN 0234-0453 • INFORMATICS AND EDUCATION • 2023 • Volume 38 No 1

DOI: 10.32517/0234-0453-2023-38-1-55-63

УЧЕБНО-МЕТОДИЧЕСКИЕ МАТЕРИАЛЫ МАСТЕР-КЛАССА «СОСТЯЗАТЕЛЬНЫЕ АТАКИ НА НЕЙРОННЫЕ СЕТИ РАСПОЗНАВАНИЯ ИЗОБРАЖЕНИЙ» ДЛЯ СТУДЕНТОВ И ШКОЛЬНИКОВ

Д. В. Пантюхин[1,2] ✉

[1] *Национальный исследовательский университет «Высшая школа экономики», г. Москва, Россия*

[2] *МИРЭА — Российский технологический университет, г. Москва, Россия*

✉ *dpantiukhin@hse.ru*

Аннотация

Проблема уязвимости нейронных сетей уже несколько лет является предметом научных исследований и экспериментов. Состязательные атаки — один из способов «обмануть» нейросеть, заставить ее принимать ошибочные классификационные решения. Сама возможность состязательной атаки кроется в особенностях машинного обучения нейросетей. В статье показано, как свойства нейронных сетей становятся источником проблем и ограничений в их использовании. Материалы соответствующих исследований автора легли в основу мастер-класса «Состязательные атаки на нейронные сети распознавания изображений».

В статье представлены учебно-методические материалы мастер-класса: теоретические основы занятия, практические материалы (в частности, описана атака на один нейрон, рассмотрен метод быстрого знака градиента для атаки на нейронную сеть), примеры экспериментов и расчетов (автор использует сверточную сеть VGG, библиотеки Torch и CleverHans), а также набор типовых ошибок обучающихся и пояснения педагога, устраняющие эти ошибки. Кроме того, в статье приведен результат эксперимента, а по указанным ссылкам доступны его полный код и примеры апробации материалов мастер-класса.

Мастер-класс предназначен как для студентов, так и для старших школьников, изучивших основы нейронных сетей и язык Python, а также может представлять практический интерес для учителей информатики, разработчиков курсов по машинному обучению и искусственному интеллекту и преподавателей вузов.

Ключевые слова: нейронная сеть, состязательная атака, метод быстрого знака градиента, гиперпараметры атаки, мастер-класс.

Для цитирования:

Пантюхин Д. В. Учебно-методические материалы мастер-класса «Состязательные атаки на нейронные сети распознавания изображений» для студентов и школьников. *Информатика и образование.* 2022;38(1):55–63. DOI: 10.32517/0234-0453-2023-38-1-55-63

EDUCATIONAL AND METHODOLOGICAL MATERIALS OF THE MASTER CLASS "ADVERSARIAL ATTACKS ON IMAGE RECOGNITION NEURAL NETWORKS" FOR STUDENTS AND SCHOOLCHILDREN

D. V. Pantiukhin[1,2] ✉

[1] *National Research University Higher School of Economics, Moscow, Russia*

[2] *MIREA — Russian Technological University, Moscow, Russia*

✉ *dpantiukhin@hse.ru*

Abstract

The problem of neural network vulnerability has been the subject of scientific research and experiments for several years. Adversarial attacks are one of the ways to “trick” a neural network, to force it to make incorrect classification decisions. The very possibility of adversarial attack lies in the peculiarities of machine learning of neural networks. The article shows how the properties of neural networks become a source of problems and limitations in their use. The materials of the corresponding researches of the author were used as a basis for the master class “Adversarial attacks on image recognition neural networks”.

ISSN 0234-0453 • ИНФОРМАТИКА И ОБРАЗОВАНИЕ • 2023 • Том 38 № 1

The article presents the educational materials of the master class: the theoretical background of the class, practical materials (in particular, the attack on a single neuron is described, the fast gradient sign method for attacking a neural network is considered), examples of experiments and calculations (the author uses the convolutional network VGG, Torch and CleverHans libraries), as well as a set of typical errors of students and the teacher's explanations of how to eliminate these errors. In addition, the result of the experiment is given in the article, and its full code and examples of approbation of the master class materials are available at the above links.

The master class is intended for both high school and university students who have learned the basics of neural networks and the Python language, and can also be of practical interest to computer science teachers, to developers of courses on machine learning and artificial intelligence as well as to university teachers.

Keywords: neural network, adversarial attack, fast gradient sign method, hyperparameters of attack, master class.

For citation:

Pantiukhin D. V. Educational and methodological materials of the master class "Adversarial attacks on image recognition neural networks" for students and schoolchildren. *Informatics and Education.* 2022;38(1):55–63. (In Russian.) DOI: 10.32517/0234-0453-2023-38-1-55-63

1. Введение

Область исследования нейронных сетей и их машинного обучения стремительно молодеет. То, что раньше считалось доступным только магистрантам и аспирантам, теперь преподается и школьникам. Появились многочисленные курсы [1], соревнования, хакатоны по машинному обучению, ориентированные на старших школьников и младших студентов, и подобные мероприятия пользуются широким спросом. Однако их организаторы при постановке задач часто обходят стороной проблемы и даже угрозы, которые возникают при использовании нейронных сетей.

Нейронные сети способны лучше человека решать множество сложных задач. Нейронные сети умеют распознавать и создавать изображения[1], писать музыку [2] и стихи [3], синтезировать голос [4], играть в игры [5] и многое другое. Казалось бы, искусственно создан полноценный интеллект, который может полностью заменить человека (как в фантастических фильмах). Однако это не так: нейронные сети «думают» не так, как человек. И в некоторых ситуациях, когда человек легко находит решение, нейронная сеть ошибается, как бы сложно написана она ни была. Более того, большинство нейронных сетей очень легко обмануть [6].

Возможность такого обмана, или атаки на нейронную сеть, представляет очень серьезную опасность. Представьте: едет автомобиль, который самостоятельно распознает дорожные знаки, а злоумышленник наклеил на знак маленькую заплатку, и нейронная сеть автомобиля при расшифровке знака ошибается. Результат — авария.

Сегодня существует много типов атак на нейронные сети. Все сети уязвимы перед ними в большей или меньшей степени [7]. Например, атака на системы распознавания изображений заключается в следующем: немного изменяя входное изображение, можно заставить нейронную сеть неверно распознавать его (относить к ложному классу).

На наш взгляд, недостаточно рассказывать о достижениях нейронных сетей, хотя они действительно очень значительны. Необходимо постоянно указывать на ограниченность их возможностей. С этой целью и был создан мастер-класс «Состязательные атаки на нейронные сети распознавания изображений» — как дополнение для курсов по машинному обучению и искусственному интеллекту, позволяющее более глубоко понимать принципы работы нейронных сетей и актуальные проблемы в этой области.

Мастер-класс предназначен как для студентов, так и для старших школьников, изучивших язык Python и основы создания и функционирования нейронных сетей. Для освоения его материалов достаточно знаний математики и информатики, полученных в средней школе (в частности, необходимо владеть понятием «производная»). Изложение ведется простым языком, без сложных математических конструкций.

Мастер-класс содержит теоретические и практические материалы по теме состязательных атак, которые доступны в интерактивной онлайн-системе Google Colab бесплатно. Мастер-класс может представлять практический интерес для учителей информатики, разработчиков курсов по машинному обучению и искусственному интеллекту и преподавателей вузов. Его положения и выводы были апробированы в многочисленных курсах автора.

Материальную основу мастер-класса составляют язык Python, интерактивная среда разработки Jupyter Notebook (файлы типа *.ipynb) и библиотеки для машинного обучения, в том числе:

- NumPy[2] — для работы с массивами;
- Torch[3] — для работы с нейронными сетями;
- torchvision[4] — для обработки изображений;
- CleverHans[5] — для проведения атак.

Облачная среда для работы с кодом Google Colaboratory[6] позволяет запускать примеры с любого компьютера, имеющего доступ в интернет, без специальной настройки. Сложность примеров подобрана так, чтобы они выполнялись быстро, что дает возможность проводить эксперименты непосредственно на уроке.

Длительность мастер-класса (без учета введения и самостоятельных экспериментов) составляет 2,5–3 часа.

---

1 *Roose K.* An A.I.-generated picture won an art prize. Artists aren't happy. *The New York Times.* Sept. 2, 2022. https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html

2 NumPy. https://numpy.org/

3 Torch. http://torch.ch/, http://pytorch.org/

4 Библиотека torchvision. PyTorch. https://pytorch.org/vision/stable/index.html

5 Исходный код CleverHans: https://github.com/cleverhans-lab/cleverhans

6 Google Colaboratory. https://colab.research.google.com/

ISSN 0234-0453 • INFORMATICS AND EDUCATION • 2023 • Volume 38 No 1

2. Теоретические основы мастер-класса: нейронные сети и особенности их обучения

Предполагается, что слушатель имеет базовые представления о нейронных сетях. Однако мы кратко поясняем основные понятия, которые потребуются для последующей работы, и напоминаем об обязательных к освоению исследованиях [8, 9, 10, 11].

Нейронная сеть представляет собой вычислительную архитектуру, которая занимается преобразованием входных данных в выходные данные по заданным правилам. Нейронная сеть состоит из слоев, соединенных между собой и выполняющих преобразования входной информации. Основные слои имеют параметры, которые могут быть изменены. При изменении параметров изменяется (часто — существенно) и результат преобразования, которое выполняет соответствующий слой. Значит, можно так подбирать параметры слоев, чтобы преобразование, выполняемое сетью целиком, соответствовало требованиям. Это и есть процесс обучения нейронной сети.

Задача обучения нейронной сети — задать такие параметры слоев, чтобы выходные данные (выходы) нейронной сети в целом при известных входных данных (входах) были как можно ближе к тем выходам, которые мы хотели бы получать. Мера близости формулируется в виде функции ошибки, и обучение нейронной сети сводится к задаче оптимизации функции ошибки по параметрам слоев. Практически все методы обучения основаны на вычислении градиента (вектора производных) функции ошибки по обучаемым параметрам, для чего используется метод обратного распространения ошибки, позволяющий пересчитывать градиент послойно, начиная с последнего слоя.

Возможность такого послойного пересчета градиента положена в основу методов автодифференцирования [12], которые реализованы в современных библиотеках машинного обучения. Компьютер может рассчитывать производные заданных функций автоматически, без явного выписывания формул.

В области обработки изображений базовой задачей является распознавание изображения [13, 14]: нейронная сеть должна вернуть класс того изображения, которое было подано на вход. Обычно такие нейронные сети имеют последний слой с числом нейронов, соответствующим числу классов, которые могут быть получены при распознавании. Тогда каждый выход можно интерпретировать как уровень уверенности сети в правильном распознавании класса, за который такой выход отвечает. При использовании функций активации типа softmax [15] сумма всех выходов будет равна 1, а сами выходы будут принимать значения из диапазона от 0 до 1. Считается, что сеть распознает такой класс, у которого самая большая уверенность. Сейчас нас не интересует внутреннее устройство нейронных сетей, важно, что они принимают на вход изображение и возвращают вектор уверенностей в распознанных классах. Обычно сети, предназначенные для распознавания изображений, — это сверточные нейронные сети.

В своих экспериментах мы будем использовать уже обученную нейронную сеть типа VGG [16], которая является представителем сверточных сетей. Это действительно хорошо обученная сеть (обучалась на миллионе изображений): она умеет верно распознавать 1000 классов изображений с высокой точностью (порядка 75 %), содержит 140 млн обучаемых параметров, обрабатывает цветные изображения размером 224 × 224 пикселя (рис. 1).

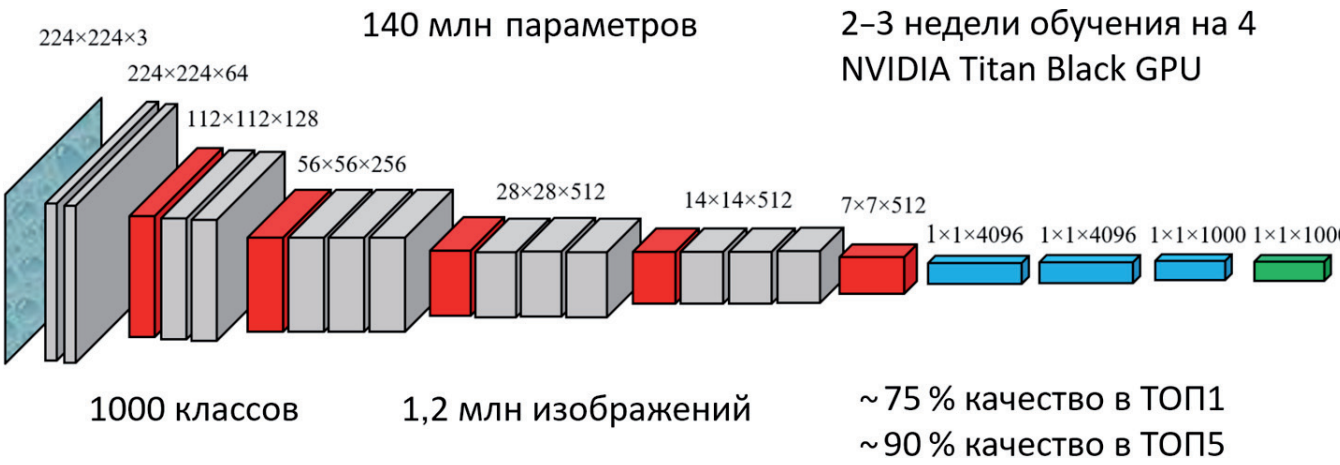


Рис. 1. Изображение архитектуры сети VGG[1]
Fig. 1. Image of VGG neural network architecture

[1] Источник оригинального изображения: https://upload.wikimedia.org/wikipedia/en/8/83/VGG_neural_network.png Информацию об архитектуре сверточной сети VGG см.: https://paperswithcode.com/paper/very-deep-convolutional-networks-for-large

ISSN 0234-0453 • ИНФОРМАТИКА И ОБРАЗОВАНИЕ • 2023 • Том 38 № 1

3.Мастер-класс: расчеты и виды атаки на нейронную сеть

3.1. Состязательные атаки на нейронные сети

В нашем эксперименте с помощью именно этой, хорошо обученной сети VGG была предпринята попытка распознать изображение (см. рис. 2, *а*). Нейронная сеть идентифицировала его отлично: это собака породы «самоед»[1], 66 % уверенности. Соответствующий класс («samoyed») был задан при обучении этой нейронной сети. Ближайшие предположения: 16 % уверенности в классе «шпиц» («pomeranian») и чуть больше 1 % — «другие породы собак». Действительно, распознавание очень уверенное и корректное.

Проверка работы этой же хорошо обученной нейронной сети со следующим изображением (см. рис. 2, б) показывает, что сеть VGG уверенно (на 100 %) распознает на нем страуса (класс «ostrich»).

Но как же так? Для нас это одна и та же картинка. Изображения на рисунке 2, *а* и на рисунке 2, *б* для глаза человека неразличимы. Но нейронная сеть распознает их по-разному. Если изучить числовое описание этих изображений, то увидим, что они отличаются на небольшую величину — возмущение (см. на рис. 2, *в* при стократном увеличении), которое незаметно глазу. Такой маленькой добавки хватило для того, чтобы обмануть нейронную сеть.

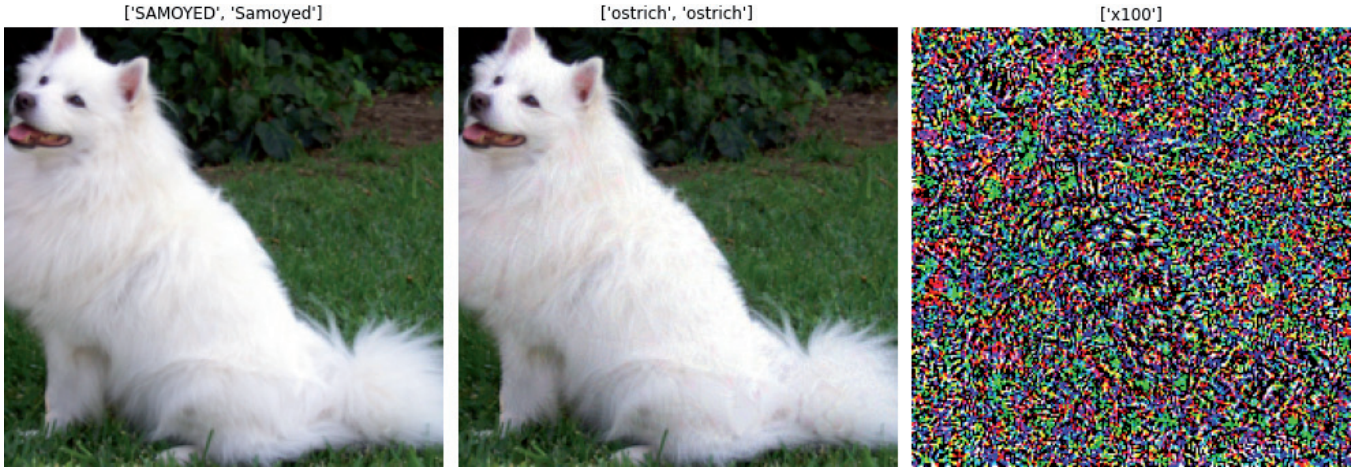
Добавка похожа на случайный шум. Если добавить на изображение слабый случайный шум, то нейронная сеть распознает ее без ошибок. Дело в том, что добавленный выше «шум» *не случайный*: он был специальным образом рассчитан так, чтобы обмануть нейронную сеть.

Процесс создания специальных добавок к входным данным нейронной сети, которые обманывают ее, получил название «состязательной атаки» на нейронную сеть (adversarial attack). Нейронные сети подвержены атакам [17, 18]!

3.2. Пример состязательной атаки на один нейрон

Как «думает» нейронная сеть? Взять пиксели, умножить их на веса, сложить произведения, пропустить сумму через функцию активации. Сделать так столько раз, сколько слоев в сети — получить выход. А что будет, если мы возьмем один пиксель и будем бесконечно увеличивать его значение (вес пикселя ненулевой)? Остальные пиксели потеряются на его фоне, от него будет получаться очень большое слагаемое, а другие пиксели дадут сравнительно маленький вклад. И только этот пиксель будет определять выход сети.

В реальности мы не можем изменять пиксель бесконечно, ведь значения пикселей ограничены величиной 255 в целочисленном представлении. Тогда можно изменять все пиксели на небольшую величину, но так, чтобы суммарные изменения выхода были большими.



а *б* *в*

Рис. 2. Изображение собаки «самоед»[2]:
а — нормальное распознавание, уверенность — 66 %;
б — измененное изображение собаки, ошибочное распознавание как «страус», уверенность — 100 %;
в — отличие изображений а и б, стократное увеличение

Fig. 2. Samoyed image:
а — correct recognition 66 %;
б — changed image, incorrect recognition as 100 % ostrich;
в — difference between a and б images with 100 scale factor

[1] Самоедская собака. *Свободная энциклопедия «Википедия»*. https://ru.wikipedia.org/wiki/Самоедская_собака
[2] Источник оригинального изображения: https://github.com/pytorch/hub/raw/master/images/dog.jpg

ISSN 0234-0453 • INFORMATICS AND EDUCATION • 2023 • Volume 38 No 1

Пусть:
x_i — входы;
δx_i — маленькие добавки для входов;
w_i — веса входов.

Тогда на примере одного нейрона без функции активации и смещений для нормальной ситуации получаем выход:

$$y(x) = w_1 \cdot x_1 + w_2 \cdot x_2 + \ldots = \sum_i (w_i \cdot x_i).$$

При атаке выход получается:

$$y_{\text{attack}}(x) = w_1 \cdot (x_1 + \delta x_1) + w_2 \cdot (x_2 + \delta x_2) + \ldots = \\ = \sum_i (w_i \cdot (x_i + \delta x_i))$$

или

$$y_{\text{attack}}(x) = y(x) + \sum_i (w_i \cdot \delta x_i).$$

Если даже каждая добавка входа δx_i маленькая, то добавка выхода $\sum_i (w_i \cdot \delta x_i)$ не обязательно должна быть маленькой. Например, если взять $\delta x_i = \varepsilon \cdot w_i$, где ε — маленькое число, добавка выхода $\varepsilon \cdot \sum_i (w_i \cdot w_i)$ при большом числе слагаемых будет отнюдь не маленькой, ведь все слагаемые здесь неотрицательные. Ее даже может хватить для того, чтобы нейрон стал возвращать другой класс.

Мы провели атаку на один нейрон и при этом рассчитали добавки входов, которые действительно могут обмануть его, в зависимости от параметров нейрона. Чем больше входов, тем больше шанс, что суммарная добавка выхода нейрона будет достаточной для обмана при маленьких добавках входов, т. е. атака сработает и будет незаметной. Это как раз та ситуация, с которой мы сталкиваемся при обработке изображений, — пикселей в изображении довольно много.

3.3. Метод быстрого знака градиента для атаки на нейронную сеть

Для более сложных сетей, состоящих из множества нейронов, мы можем посчитать, насколько необходимо изменить каждый пиксель, чтобы это изменение максимально искажало выход. Для этого следует найти градиент (производную) функции ошибки по соответствующему пикселю. При обучении сети мы, изменяя веса, минимизировали функцию ошибки по параметрам, а теперь будем оптимизировать ошибку, изменяя вход.

В рамках задач распознавания атаки могут быть *направленными*, когда требуется получить какой-то конкретный класс при распознавании, или *ненаправленными*, когда неважно, какой класс получится при распознавании, лишь бы не настоящий.

Простой метод атаки на нейронную сеть — метод быстрого знака градиента[1] [18] (Fast Gradient Sign Method, FGSM).

Предположим, что у нас есть обученная нейронная сеть (например, она распознает изображения). Возьмем изображение, для которого хотим провести атаку. Посчитаем градиент функции ошибки по всем пикселям изображения (нам нужен только знак градиента). Изменим пиксели так, чтобы максимизировать функцию ошибки. Если при обучении нейронной сети для минимизации ошибки мы использовали метод градиентного спуска, то теперь будем максимизировать ошибку по значениям пикселей. Чтобы изменения были маленькими (незаметными), умножим прибавляемый градиент на малое число ε. При необходимости повторим процедуру. Важно не перестараться, чтобы общие изменения были маленькими.

В рамках задач классификации, когда нейронная сеть возвращает нам уровни уверенности в классах, процедура упрощается: пусть F_i — i-й выход, который отвечает за уровень уверенности в классе i. Для направленной атаки есть целевой класс G, который мы хотели бы получить на выходе. Значит, необходимо увеличивать уверенность F_G для этого целевого класса. Поскольку сумма уверенностей по всем классам равна 1 (обычно для этого используют softmax), сумма уверенностей для всех остальных классов, кроме целевого, будет снижаться. В какой-то момент времени уверенность для целевого класса станет больше, чем для настоящего класса, и сеть начнет возвращать целевой класс. Пиксели изменяются согласно правилу:

$$x_i = x_i + \varepsilon \cdot \text{sign} \frac{\partial F_G}{\partial x_i}. \quad (1)$$

Для ненаправленной атаки, когда неважно, какой класс будет возвращен, лишь бы не настоящий R, уверенность F_R для настоящего класса R необходимо снижать. Сумма уверенностей для всех остальных классов при этом будет увеличиваться. В какой-то момент времени уверенность для настоящего класса станет меньше, чем для одного из других классов, и сеть начнет возвращать другой класс. Пиксели изменяются согласно правилу:

$$x_i = x_i - \varepsilon \cdot \text{sign} \frac{\partial F_R}{\partial x_i}. \quad (2)$$

Технически реализация направленной и ненаправленной атаки отличаются незначительно.

4. Пример реализации состязательной атаки на сети распознавания изображений

В мастер-классе приведен пример атаки на нейронную сеть, которая распознает изображения. Для этого воспользуемся библиотекой Torch, а реализацию метода FGSM возьмем из библиотеки CleverHans. Мы будем использовать уже обученную нейронную сеть типа VGG[2]. Сети такого типа относятся к самым эффективным для распознавания изображений.

[1] *Cheng*. Adversarial Attack and Defense on Neural Networks in PyTorch. Do neural networks really learn everything? *Towards Data Science*. Sept. 8, 2020. https://towardsdatascience.com/adversarial-attack-and-defense-on-neural-networks-in-pytorch-82b5bcd9171

[2] https://colab.research.google.com/github/pytorch/pytorch.github.io/blob/master/assets/hub/pytorch_vision_vgg.ipynb

ISSN 0234-0453 • ИНФОРМАТИКА И ОБРАЗОВАНИЕ • 2023 • Том 38 № 1

Общий подход к реализации атаки:

- установить и подключить все необходимые библиотеки;
- скачать, подключить, инициализировать обученную нейронную сеть для распознавания изображений;
- подготовить атакуемое изображение, в том числе масштабировать его до размеров, поддерживаемых сетью распознавания (224 × 224 пикселя в примерах);
- проверить, что это изображение распознается верно, для чего подать изображение на обученную сеть распознавания, вернуть уровни уверенности в распознанных классах, выбрать наиболее уверенно распознанный класс, сравнить его с настоящим классом;
- выполнить атаку, для чего задать гиперпараметры атаки (направленная или нет, целевой класс, сила атаки ε, число шагов атаки), в цикле по числу шагов атаки выполнить расчет градиентов и использовать их для обновления пикселей согласно правилам (1) или (2);
- проверить, что атака выполнена успешно, для чего посчитать класс, возвращаемый нейронной сетью для атакованного изображения, проверить, что возвращаемый класс не совпадает с настоящим и (в случае направленной атаки) совпадает с целевым классом;
- проверить визуально, что атака незаметна или малозаметна для человека.

Результаты эксперимента приведены на рисунке 2. Полный код эксперимента доступен по ссылке[1]. В этом эксперименте атака прошла успешно и действительно незаметна. Использовались следующие гиперпараметры:

- сила атаки — 0,01;
- число шагов атаки — 20;
- направленная атака;
- целевой класс — № 9 «ostrich».

Слушателям предлагается самостоятельно провести исследования работы алгоритма атаки в различных ситуациях, в том числе:

- менять силу атаки (число ε);
- менять количество шагов атаки;
- менять нейронную сеть, подвергаемую атаке;
- менять атакуемые изображения, в том числе их исходный размер;
- менять целевой класс в направленной атаке.

В рамках этих экспериментов слушатель может самостоятельно убедиться в том, что:

- нейронные сети подвержены атакам, если только не предпринимались специальные действия по защите от атак;
- многие атаки сработают, независимо от атакуемого изображения;
- возможность атаки не гарантирована, могут быть случаи, когда атака не сработает;
- гиперпараметры (сила атаки и число шагов атаки) значительно влияют как на успешность атаки, так и на ее незаметность;
- уменьшение количества пикселей атакуемого изображения обычно ведет не только к затруднению атаки, но и к ухудшению качества распознавания не атакованного изображения (что опять же не гарантировано).

5. Апробация материалов

5.1. Проведение мастер-класса

Мастер-класс проводился несколько раз:

- в рамках мероприятий «Российского движения школьников»: доклад о курсе для школьников по машинному обучению и нейронным сетям, г. Москва, 2021 г., ~100 участников[2];
- в рамках семинара по проблемам информационной безопасности на базе МИВлГУ, тема семинара «Состязательные атаки на нейронные сети распознавания изображений», г. Муром, 2022 год, ~40 участников[3];
- в рамках мероприятий Детского технопарка «Альтаир» РТУ МИРЭА: онлайн-лекция «Соревновательные атаки на нейронные сети классификации изображений», 2022 год, ~100 участников[4];
- в рамках мероприятий акселератора Leader-ID: лекция «Атаки на искусственный интеллект», 2022 год, ~30 участников[5].

5.2. Использование материалов мастер-класса

Материалы мастер-класса вошли в состав учебных курсов, читаемых автором, в том числе:

- курс «Машинное обучение для школьников» (X-XI классы) для лаборатории «Наносемантика»[6] и ГАОУ ДПО «ТемоЦентр», в том числе вебинары для учителей информатики по разработанным материалам (2021 год);
- программа повышения квалификации педагогических работников образовательных организаций высшего образования в сфере

[1] https://drive.google.com/file/d/1QFzfSKP0kU5oijW0myvQffuM88ZO1QxF/view?usp=sharing

[2] Курсы машинного обучения и нейросетевых технологий для учащихся X-XI классов. Вторая часть. *VK Видео. Видеозаписи Российского движения школьников. #Объяснитенормально | «Наносемантика»*. https://vk.com/video/@skm_rus?z=video-122623791_456245726%2Fclub122623791%2Fpl_-122623791_-2

[3] Семинар по проблемам информационной безопасности. *Муромский институт*. 13 апреля 2022 года. https://www.mivlgu.ru/news/2022-04/obrazovanie/seminar-po-problemam-informacionnoy-bezopasnosti

[4] Онлайн-лекция «Соревновательные атаки на нейронные сети классификации изображений» от Детского технопарка «Альтаир» РТУ МИРЭА. *Приемная комиссия МИРЭА — Российского технологического университета*. 15 октября 2022 года. https://priem.mirea.ru/event?event_id=1704

[5] Лекция «Атаки на искусственный интеллект». *Leader-ID*. 3 декабря 2022 года. https://leader-id.ru/events/350432

[6] Наносемантика. Ведущие разработчики речевых технологий на базе искусственного интеллекта. https://nanosemantics.ai/

ISSN 0234-0453 • INFORMATICS AND EDUCATION • 2023 • Volume 38 No 1

Таблица / Table

Типовые ошибки и разъяснения по ним
Typical errors when testing on the materials of the master class and explanations

Вопросы и ошибки	Разъяснения
Отвечая на вопрос «Что меняют внутри нейронной сети при совершении атаки?», часто указывают параметры нейронной сети (веса, смещения и др.)	Атака не меняет саму нейронную сеть, меняются только входы в такую сеть (например, изображение). Ошибка связана с похожестью процедуры атаки на процедуру обучения, когда параметры действительно меняются
Отвечая на вопрос о размере массива, описывающего рассчитываемый градиент при совершении атаки, часто пытаются учитывать размер самой сети	Градиент считается только по пикселям изображений. **Параметры нейронной сети не меняются, считать по ним градиент не нужно. Ошибка связана с похожестью процедуры атаки на процедуру обучения, когда параметры действительно меняются**
Забывают, что цветные изображения описываются не одной матрицей, а несколькими (по числу каналов цвета)	При представлении изображений в цветовой модели RGB размер массива градиента в три раза больше числа пикселей изображения, ведь каждый пиксель описывается тремя числами
Отвечая на вопрос о различии процедур направленной и ненаправленной атаки методом FGSM, часто указывают только разный знак градиента, забывая про другие отличия	Процедуры направленной и ненаправленной атаки различаются не только знаком, но и тем, какая переменная оптимизируется. Для направленной атаки максимизируется уверенность целевого класса (знак «+»), для ненаправленной атаки минимизируется уверенность настоящего класса (знак «–»). Это разные переменные
Считают, что большие значения силы атаки (число ε) или числа шагов атаки приводят к большему успеху атаки	Это не так. Атака успешна не только в том случае, если изменился выход сети, но и если она незаметна (малозаметна). Поэтому успешность атаки — это компромисс между достаточностью изменения выхода сети и незаметностью. Увеличение указанных гиперпараметров ведет к понижению незаметности. Следовательно, нельзя утверждать, что большие значения ведут к большему успеху атаки. Кроме того, заранее неизвестно, что выгоднее: использовать много шагов с маленькой силой атаки или несколько шагов с большой силой атаки. Это решается в результате эксперимента: для разных сетей и для разных изображений ответ может быть различным
Ошибки реализации: • измеряют массивы в неправильном порядке; • забывают объявлять массив изображения типом данных, который поддерживает автодифференцирование; • используют неподходящие версии библиотек	Следует тщательно проверять требования к версиям используемых библиотек, особенно при использовании предустановленных библиотек в Google Colab. Нужно напоминать обучающимся о необходимости тщательно изучать документацию используемых библиотек

искусственного интеллекта «Нейронные сети анализа изображений и аудиоинформации», РТУ МИРЭА, 2022 год, ~50 слушателей;
- программы обучения РТУ МИРЭА[1]:
  – «Анализ аудио-, видеоинформации в системах безопасности», 4-й курс бакалавриата, семестровый курс, ~50 слушателей;
  – факультатив «Введение в нейронные сети», годовой, для всех студентов, ~200 слушателей;
  – «Библиотеки машинного обучения», 4-й курс бакалавриата, семестровый, ~50 слушателей;
  – «Мониторинг безопасности и киберугроз», 1-й курс магистратуры, семестровый, ~50 слушателей;
- курсы НИУ ВШЭ[2]:
  – майнор (курс по выбору) «Нейросетевые технологии»[3], 2-3-й курсы бакалавриата, двухгодичный, для всех образовательных программ, ~200 слушателей;
  – научно-исследовательский семинар «Нейросетевые технологии», 1-й курс бакалавриата, ~70 слушателей;
  – «Проектный семинар по информационной безопасности», 1-й курс магистратуры, ~30 слушателей.

Кроме того, материалы мастер-класса широко используются студентами при выполнении курсовых и проектных работ.

5.3. Типовые ошибки обучающихся при тестировании по итогам мастер-класса

Результаты анализа ошибок, допускаемых при тестировании по итогам мастер-класса, а также пояснения педагога, которые устраняют заблуждения и просчеты, приведены в таблице.

[1] Программа обучения в магистратуре РТУ МИРЭА «10.04.01 Информационная безопасность». МИРЭА — *Российский технологический университет*. https://www.mirea.ru/education/the-institutes-and-faculties/institut-kiberbezopasnosti-i-tsifrovykh-tekhnologiy/training-programs/magistratura/10-04-01-information-security/

[2] Учебные курсы 2022/2023 учебные годы. *Национальный исследовательский университет «Высшая школа экономики»*. https://www.hse.ru/staff/PantiukhinDmitry#teaching

[3] Майнор «Нейросетевые технологии». *Национальный исследовательский университет «Высшая школа экономики»*. Выбор траектории обучения. https://electives.hse.ru/neuronet/

ISSN 0234-0453 • ИНФОРМАТИКА И ОБРАЗОВАНИЕ • 2023 • Том 38 № 1

5.4. Итоги мастер-класса и его перспективы

Дополнительные материалы к мастер-классу[1] раскрывают ряд смежных тем. Кратко описаны другие подходы к атакам, например малопиксельные, когда изменяется только несколько пикселей изображения [19, 20]. Показаны возможности и опасности проведения атак в физическом мире, когда атакующим изменениям подвергаются не файлы изображений, а сами фотографируемые объекты [21, 22]. Представлены подходы к совершению атак «черного ящика», когда сама атакуемая нейронная сеть неизвестна, но имеется доступ к ее входам и выходам [23]. Упомянуты другие виды атак, как, например, «отравление данных»[2]. Даются краткие сведения о возможностях защиты от состязательных атак [7], но также показано, что не гарантированы ни полный успех атак, ни полная защита от них. Сегодня постоянно появляются все новые и новые способы атак и их предотвращения. Наконец, показана связь между подходами атак и генеративно-состязательными сетями, предназначенными для создания новых правдоподобных образов[3].

В планах автора — расширение материалов мастер-класса, как в части добавления новых видов атак для экспериментов, так и в части исследований по защите от таких атак.

Благодарности

Автор выражает глубокую благодарность генеральному директору ООО «Лаборатория Наносемантика» Станиславу Игоревичу Ашманову, при поддержке которого были созданы упомянутый курс машинного обучения для школьников и данный мастер-класс, а также всем студентам, которые вычитывали материалы мастер-класса, коллегам, которые давали дельные советы, и своему учителю Александру Ивановичу Галушкину[4] за то, что научил обучать нейронные сети.

Acknowledgments

The author expresses his sincere gratitude to Stanislav Igorevich Ashmanov, the General Director of Nanosemantics Laboratory LLC, with whose support the mentioned machine learning course for schoolchildren and this master class were created. The author is also grateful to all students who proofread the materials of the master class, to the colleagues who gave useful advice and to his teacher Alexander Galushkin[4] for teaching to train neural networks.

Список источников / References

1. *Marques L. S., Gresse von Wangenheim C., Hauck J. C. R.* Teaching machine learning in school: A systematic mapping of the state of the art. *Informatics in Education*. 2020;19(2):283–321. DOI: 10.15388/infedu.2020.14

2. *Briot J.-P., Hadjeres G., Pachet F.-D.* Deep learning techniques for music generation. Cham, Springer; 2020. 284 p. DOI: 10.1007/978-3-319-70163-9

3. *Ormazabal A., Artetxe M., Agirrrezabal M., Soroa A., Agirre E.* PoeLM: A meter-and rhyme-controllable language model for unsupervised poetry generation. *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi, Association for Computational Linguistics; 2022:3655–3670. DOI: 10.48550/arXiv.2205.12206

4. *Пантюхин Д. В.* Нейронные сети синтеза речи голосовых помощников и поющих автоматов. *Речевые технологии*. 2021;(3-4):3–16. EDN: PEDILK. DOI: 10.58633/2305-8129_2021_3-4_3

[*Pantiukhin D. V.* Neural networks for speech synthesis of voice assistants and singing machines. *Speech Technology*. 2021;(3-4):3–16. (In Russian.) EDN: PEDILK. DOI: 10.58633/2305-8129_2021_3-4_3]

5. *Zhu J., Villareale J., Javvaji N., Risi S., Löwe M., Weigelt R., Harteveld C.* Player-AI interaction: What neural network games reveal about AI as play. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. New York, Association for Computing Machinery; 2021:1–17. DOI: 10.1145/3411764.3445307

6. *Long T., Gao Q., Xu L., Zhou Z.* A survey on adversarial attacks in computer vision: Taxonomy, visualization and future directions. *Computers and Security*. 2022;121:102847. DOI: 10.1016/j.cose.2022.102847

7. *Chakraborty A., Alam M., Dey V., Chattopadhyay A., Mukhopadhyay D.* A survey on adversarial attacks and defences. *CAAI Transactions on Intelligence Technology*. 2021;6(1):25–45. DOI: 10.1049/cit2.12028

8. *Николенко С., Кадурин А., Архангельская Е.* Глубокое обучение. Погружение в мир нейронных сетей. СПб.: Питер; 2018. 480 с.

[*Nikolenko S., Kadurin A., Arkhangelskaya E.* Deep learning. Immersion into the world of neural networks. Saint Petersburg, Piter; 2018. 480 p. (In Russian.)]

9. *Гудфеллоу Я., Бенджио И., Курвилль А.* Глубокое обучение. М.: ДМК-Пресс; 2018. 652 с.

[*Goodfellow I., Bengio Y., Courville A.* Deep learning. Moscow, DMK-Press; 2018. 652 p. (In Russian.)]

10. *Джулли А., Пал С.* Библиотека Keras — инструмент глубокого обучения. Реализация нейронных сетей с помощью библиотек Theano и TensorFlow. М.: ДМК-Пресс; 2017. 296 с.

[*Gulli A., Pal S.* Deep Learning with Keras. Implement neural networks with Keras on Theano and TensorFlow. Moscow, DMK-Press; 2017. 296 p. (In Russian.)]

11. *Пойнтер Я.* Программируем с PyTorch: Создание приложений глубокого обучения. СПб.: Питер; 2020. 256 с.

[*Pointer I.* Programming PyTorch for Deep Learning: Creating and Deploying Deep Learning Applications. Saint Petersburg, Piter; 2020. 256 p. (In Russian.)]

12. *Ketkar N., Moolayil J.* Deep learning with Python. Learn best practices of deep learning models with PyTorch. Berkeley, Apress; 2021. 306 p.

13. *Khan A., Sohail A., Zahoora U., Qureshi A. S.* A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*. 2020;53(8):5455–5516. DOI: 10.48550/arXiv.1901.06032

14. *Hussain M., Bird J. J., Faria D. R.* A study on CNN transfer learning for image classification. *Advances in Computational Intelligence Systems. UKCI 2018. Advances in Intelligent Systems and Computing*. Cham, Springer; 2019;840:191–202. DOI: 10.1007/978-3-319-97982-3_16

15. *Gao B., Pavel L.* On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint*. 2017:1–10. DOI: 10.48550/arXiv.1704.00805

16. *Simonyan K., Zisserman A.* Very deep convolutional networks for large-scale image recognition. *arXiv preprint*. 2015:1–14. DOI: 10.48550/arXiv.1409.1556

---

1 Лекция «Атаки на искусственный интеллект». *Leader-ID*. 3 декабря 2022 года. https://leader-id.ru/events/350432

2 Data Poisoning. Papers *With Code*. https://paperswithcode.com/task/data-poisoning

3 *Brownlee J.* A Gentle Introduction to Generative Adversarial Networks (GANs). *Machine Learning Mastery*. June 17, 2019. https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/

4 Александр Иванович Галушкин. *Свободная энциклопедия «Википедия»*. https://ru.wikipedia.org/wiki/Галушкин,_Александр_Иванович_(учёный)

ISSN 0234-0453 • INFORMATICS AND EDUCATION • 2023 • Volume 38 No 1

17. *Kianpour M., Wen S.-F.* Timing attacks on machine learning: state of the art. *Advances in Intelligent Systems and Computing*. 2020;1037:111–125. DOI: 10.1007/978-3-030-29516-5_10

18. *Goodfellow I. J., Shlens J., Szegedy Ch.* Explaining and harnessing adversarial examples. *arXiv preprint*. 2015:1–11. DOI: 10.48550/arXiv.1412.6572

19. *Carlini N., Wagner D.* Towards evaluating the robustness of neural networks. *2017 IEEE Symposium on Security and Privacy*. San Jose, 2017:39–57. DOI: 10.1109/SP.2017.49

20. *Wiyatno R., Xu A.* Maximal Jacobian-based saliency map attack. *arXiv preprint*. 2018:1–5. DOI: 10.48550/arXiv.1808.07945

21. *Eykholt K., Evtimov I., Fernandes E., Li B., Rahmati A., Xiao Ch., Prakash A., Kohno T., Song D.* Robust physical-world attacks on deep learning visual classification. *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*. 2018:1625–1634. DOI: 10.1109/CVPR.2018.00175

22. *Kurakin A., Goodfellow I. J., Bengio S.* Adversarial examples in the physical world. *Artificial intelligence safety and security*. New York, Chapman and Hall/CRC; 2018:99–112. DOI: DOI:10.1201/9781351251389-8 Available at: https://arxiv.org/pdf/1607.02533.pdf

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

Информация об авторе

Пантюхин Дмитрий Валерьевич, старший преподаватель департамента программной инженерии, факультет компьютерных наук, Национальный исследовательский университет «Высшая школа экономики», г. Москва, Россия; старший преподаватель кафедры «Интеллектуальные системы информационной безопасности», Институт кибербезопасности и цифровых технологий, МИРЭА — Российский технологический университет, г. Москва, Россия; *ORCID:* https://orcid.org/0000-0003-2088-0474; *e-mail:* dpantiukhin@hse.ru

Information about the author

Dmitry V. Pantiukhin, Senior Lecturer at School of Software Engineering, Faculty of Computer Science, National Research University Higher School of Economics, Moscow, Russia; Senior Lecturer at the Department of Intelligent Systems for Information Security, Institute for Cybersecurity and Digital Technologies, MIREA — Russian Technological University, Moscow, Russia; *ORCID:* https://orcid.org/0000-0003-2088-0474; *e-mail:* dpantiukhin@hse.ru

Поступила в редакцию / Received: 10.12.22.
Поступила после рецензирования / Revised: 30.01.23.
Принята к печати / Accepted: 31.01.23.