УДК 004.83
DOI: 10.26102/2310-6018/2024.45.2.005

Проблема компрометации системы распознавания изображений путем целенаправленной фальсификации обучающего множества

А.А. Хмелёва✉, Р.Ю. Демина, И.М. Ажмухамедов

Астраханский государственный университет им. В.Н. Татищева, Астрахань, Российская Федерация

Резюме. Работа посвящена проблеме безопасности систем распознавания изображений, основанных на использовании нейронных сетей. Подобные системы применяются в различных областях и крайне важно обеспечить их безопасность от атак, направленных на методы искусственного интеллекта. Рассмотрены сверточная нейронная сеть ResNet18, проверочное множество ImageNet для распознавания объектов на изображении и отнесения его к классу и состязательные атаки, которые направлены на изменение изображения, обрабатываемые данной нейронной сетью. Сверточные нейронные сети детектируют и сегментируют объекты, которые находятся на изображениях. Атака совершалась на этапе детектирования для того, чтобы не распознавалось присутствие объектов на изображении, а также на этапе сегментации, измененное изображение относило распознанный объект к другому классу. Реализована серия экспериментов, которая показала, как состязательная атака изменяет разные изображения. Для этого взяты изображения с животными и на них совершена состязательная атака, анализ результатов позволил определить количество итераций, необходимых для совершения успешной атаки. Также проведено сравнение исходных изображений с их модифицированными в ходе атаки версиями.

Ключевые слова: нейронные сети, атаки на нейронные сети, состязательные атаки, ResNet18, матрица превращений.

Для цитирования: Хмелёва А.А., Демина Р.Ю., Ажмухамедов И.М. Проблема компрометации системы распознавания изображений путем целенаправленной фальсификации обучающего множества. *Моделирование, оптимизация и информационные технологии.* 2024;12(2). URL: https://moitvivt.ru/ru/journal/pdf?id=1535 DOI: 10.26102/2310-6018/2024.45.2.005

The problem of compromising the image recognition system by purposefully falsifying the training set

A.A. Khmeleva✉, R.Y. Demina, I.M. Azhmukhamedov

Astrakhan State University named after V.N. Tatishchev, Astrakhan, the Russian Federation

Abstract. This work is devoted to the problem of the security of image recognition systems based on the use of neural networks. Such systems are used in various fields and it is extremely important to ensure their safety from attacks aimed at artificial intelligence methods. The convolutional neural network ResNet18, the ImageNet verification set for recognizing objects in an image and classifying it to a class, and adversarial attacks aimed at changing the image processed by this neural network are considered. Convolutional neural networks detect and segment the objects that are in the images. The attack was carried out at the detection stage in order not to recognize the presence of objects in the image, as well as at the segmentation stage, the modified image attributed the recognized object to another class. A series of experiments was implemented that showed how an adversarial attack changes different images. To do this, images with animals were taken and an adversarial attack was carried out on them, the analysis of their results allowed us to determine the number of iterations necessary to make a successful attack. The original images were also compared with their versions modified during the attack.

Keywords: neural networks, attacks on neural networks, adversarial attacks, ResNet18, transformation matrix.

For citation: Khmeleva A.A., Demina R.Yu., Azhmukhamedov I.M. The problem of compromising the image recognition system by purposefully falsifying the training set. *Modeling, Optimization and Information Technology*. 2024;12(2). URL: https://moitvivt.ru/ru/journal/pdf?id=1535 DOI: 10.26102/2310-6018/2024.45.2.005 (In Russ.).

Введение

Методы машинного обучения нашли свое применение в большом числе отраслей науки и техники: распознавание лиц в системах контроля и управления доступа (СКУД-ах), прогнозирование погоды и курса валюты и т. п. Специалисты из самых разных областей (информационная безопасность, энергетика, здравоохранение, фармацевтика, авиация и т. д.) [1] в той или иной мере полагаются на результаты работы алгоритмов искусственного интеллекта (Рисунок 1). Кроме того, наблюдается тенденция постоянного увеличения доли задач, решаемых с помощью методов машинного обучения, в различных областях. Так, например, за 15 лет (с 2005 по 2020 гг.) количество случаев внедрения методов искусственного интеллекта (ИИ) в медицинские процессы выросло почти в 62 раза [2].

Рисунок 1 – Динамика использования ИИ в российских компаниях
Figure 1 – Dynamics of the use of AI in Russian companies

Одной из наиболее часто решаемых методами ИИ задач является распознавание лиц на изображениях [3]. Рассмотрим данную задачу более детально на примере системы контроля и управления доступом. Эффективность таких систем во многом зависит от качества распознавания лиц сотрудников предприятия [4]. СКУД сначала распознает лицо человека, который пытается попасть в контролируемую зону, проверяет наличие разрешений у данного субъекта, после чего предоставляет допуск либо отказывает в нем [5]. В случае если СКУД начнет принимать неверные решения, то есть воспринимать незнакомцев как сотрудников организации (ошибка 1 рода), то возникнет ситуация, когда в периметре контролируемой зоны будут перемещается злоумышленники. Если

СКУД перестанет «узнавать своих» (ошибка 2 рода), то сотрудники не смогут своевременно попадать на свои рабочие места и приступать к выполнению профессиональных обязанностей, что, в свою очередь, может негативно сказаться на качестве их деятельности. Соответственно, чрезвычайно важно обеспечить корректное функционирование подобных систем, сведя к минимуму количество вышеупомянутых ошибок.

Первым этапом распознавания является выделения отдельных объектов на изображении. Для решения этой задачи используются, как правило, сверточные нейронные сети [6]. Сети могут определять на изображениях объекты и эффективно решают задачи многоклассовой классификации [7]. Нейронные сети детектируют и сегментируют объекты на изображениях. В ходе детектирования определяется присутствие объектов и их положение на изображении, а на этапе сегментации происходит отнесение выделенного объекта к какому-либо классу. Но для того, чтобы нейронная сеть могла классифицировать изображение, ее нужно обучить. От качества обучения будет зависеть корректность работы нейронной сети. В случае, если для обучения будут использоваться данные из открытых источников, разработчик должен быть уверен, правильно ли алгоритм воспринимает изображение. Злоумышленник в ходе проведения так называемых состязательных атак [8] может целенаправленно исказить данные, используемые для обучения нейронной сети, и добиться некорректной работы модели.

При этом одной из самых распространенных сетевых архитектур является ResNet. В данной статье для классификации изображений будет использоваться частный случай такой сверточной нейронной сети глубиной 18 слоев (ResNet18) [9]. Поэтому целью данной статьи стал анализ стойкости указанной нейронной сети к состязательной атаке.

Типы атак на системы распознавания, основанные на технологии нейронных сетей

Злоумышленник может осознанно повлиять на работу алгоритма распознавания объектов путем замены или модификации файлов, хранящихся в обучающем множестве. При этом необходимо отметить, что визуально изображения не претерпевают существенных изменений, заметных для человеческого глаза. Данный нюанс приводит к сложности выявлении факта вмешательства в систему распознавания.

Существует два основных типа атак на подобные системы:

1. Атаки на этапе детектирования

Рассматриваемая нейронная сеть ResNet18 может корректно обнаруживать лица людей на фотографиях. Модель корректно выполняет поставленную задачу, но если злоумышленник к изображению добавит небольшой шум, то алгоритм уже не будет находить лицо на изображении. На Рисунке 2 можно увидеть, что до атаки нейронная сеть могла обнаружить лицо человека на изображении, но после того, как была совершена атака, лицо уже не распознается, хотя внешне изображения не отличаются, т. е. возникает ошибка 2 рода.

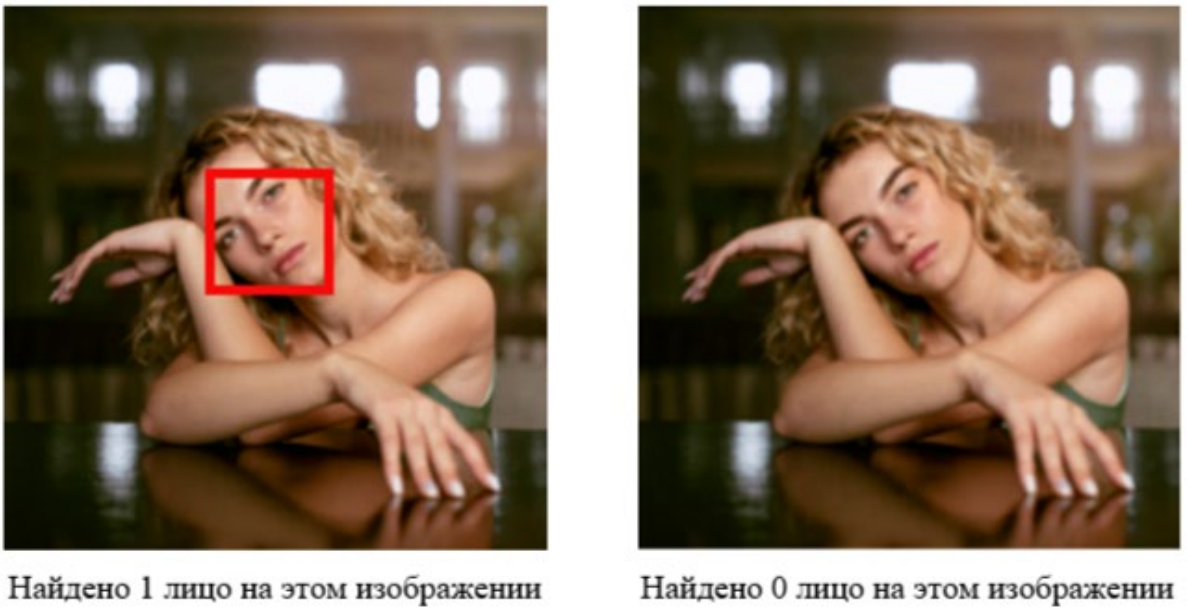


Рисунок 2 – Пример работы атаки на этапе детектирования
Figure 2 – An example of how an attack works at the detection stage

В данном случае для успешного проведения атаки к изображению добавлялся шум (отклонение) на протяжении 30 эпох (итераций). Для того, чтобы изменения были незаметными, интенсивность шума не превышала определенных границ.

2. Атака на этапе сегментирования

Особенность атаки в том, что объект на изображении распознается иначе, система относит его к другому классу. Данная атака также относится к классу состязательных.

Рассматриваемая нейронная сеть ResNet18 может корректно определять животных на фотографиях. Чтобы объекты (животные) распознавались иначе (например, медведь определялся как петух), предлагается минимизировать расстояния между состязательным примером и первоначальным изображением, которое изменяется при одновременном смешении прогноза в сторону заданного результата.

Для того, чтобы получить состязательный образец, злоумышленник реализует следующий алгоритм:

1. Загрузить исходное изображение $X\{x_i\}$, которое представляет собой множество значений пикселей по RGB каналам.

2. Нормализовать значения пикселей. Для этого значение каждого RGB канала переводится из значений в диапазоне [0,255] в значения в диапазоне [0,1] по формуле (1):

$$y_i = (x_i - \min(x)) / (\max(x) - \min(x)). \quad (1)$$

3. Сформировать тензор – массив отклонений r_i, которые лежат в определенном диапазоне и в дальнейшем будут добавлены к входному изображению X.

4. Сформировать модифицированное изображение $Y\{y_i\}$ путем добавления тензор к входному изображению $y_i = x_i + r_i$ (Рисунок 3).

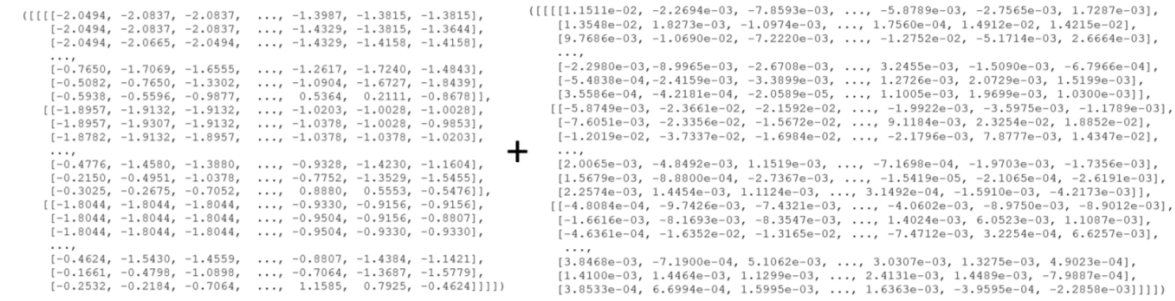


Рисунок 3 – Формирование состязательного образца
Figure 3 – The formation of an adversarial sample

5. Определить класс, к которому относится модифицированное изображение с помощью нейронной сети.

6. Если модифицированное изображение начало распознаваться некорректно, то процесс изменения изображения останавливается. Если изображение продолжает распознаваться правильно, то необходимо вычислить с использованием градиентного спуска вектор шума (отклонения), который необходимо наложить на текущее изображение и вернуться к шагу 4. Это необходимо для того, чтобы классификация изображения выполнялась так, как указал злоумышленник.

Результат атаки показан на Рисунке 4. Внешне изображения не отличаются, но изображение слева нейронной сетью классифицируется верно, а изображение справа после совершения атаки классифицируется нейронной сетью как петух. Хотя визуально изображения не отличаются друг от друга.

Рисунок 4 – Пример состязательной атаки
Figure 4 – An example of an adversarial attack

Рассмотрим более детально процесс проведения состязательной атаки на систему распознавания животных.

Пример состязательной атаки

Для примера были взяты предварительно обученная нейронная сеть ResNet18 [10] и проверочное множество ImageNet. После запуска процесса распознавания нейронная сеть корректно определяет животных, например, аксолотля (Рисунок 5).

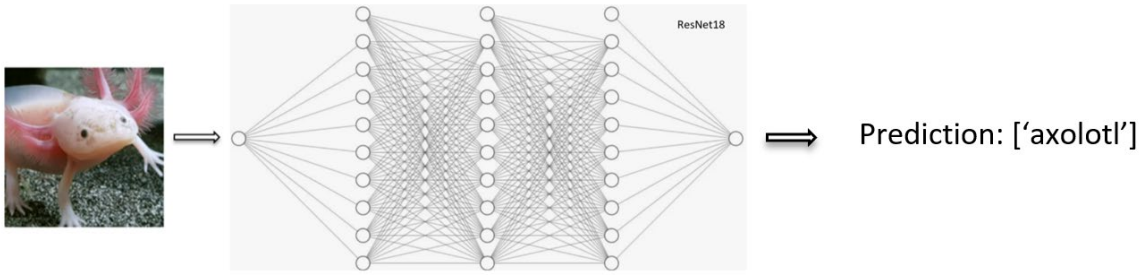


Рисунок 5 – Результат корректной работы нейронной сети
Figure 5 – The result of the correct operation of the neural network

Далее было проведено синтезирование состязательного образца, который будет нейронной сетью распознаваться как кот. Для этого необходимо запустить трехступенчатый процесс:

− преобразование изображения;
− применение отклонения;
− обратное преобразование изображения с отклонением.

После выполнения шага 3 получается модифицированное изображение «состязательный образец». Для реализации атаки необходимо выполнить следующие действия.

1. Задать значение класса, в которое будет преобразовано изображение, и значение, которое ограничивает отклонение. Это необходимо, чтобы внешне изменение не было заметно.

2. Загрузить исходное изображение.

3. Определить класс исходного изображения с помощью нейронной сети.

4. Реализовать обучающий цикл для отклонения r. Здесь формируется состязательный образец алгоритмом, который был описан выше.

5. Определить класса, к которому относится модифицированное изображение. Результат работы представлен на Рисунке 6.

```
Класс преобразования: ['n02123159', 'tiger_cat']
Величина потерь: 35.28 -- Класс: ['n01632777', 'axolotl']
Величина потерь: 6.86 -- Класс: ['n01632777', 'axolotl']
Величина потерь: 2.44 -- Класс: ['n01632777', 'axolotl']
Величина потерь: 1.06 -- Класс: ['n02123159', 'tiger_cat']
Величина потерь: 0.62 -- Класс: ['n02123159', 'tiger_cat']
Величина потерь: 0.44 -- Класс: ['n02123159', 'tiger_cat']
Старое предсказание: ['n01632777', 'axolotl']
Новое предсказание: ['n02123159', 'tiger_cat']
```

Рисунок 6 – Результат работы состязательной атаки
Figure 6 – The result of an adversarial attack

Для того, чтобы на изображении с аксолотлем модель находила кота, нейронной сети потребовалось обучение в процессе 20 эпох (Рисунок 7).

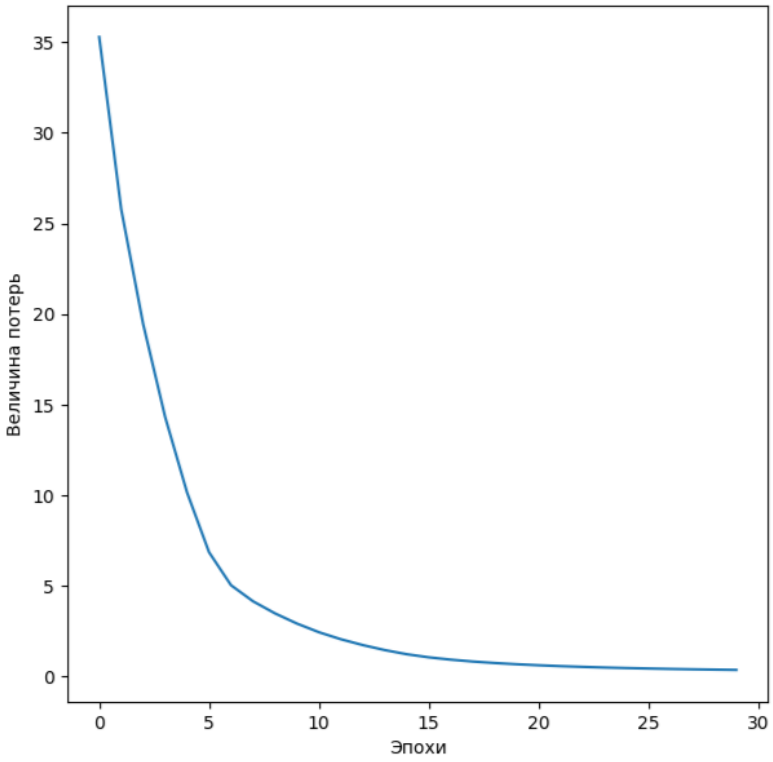


Рисунок 7 – График изменения изображения
Figure 7 – Graph of image changes

На Рисунке 8 слева показано изображение до искажения, а справа – после искажения. Как видно из рисунка, внешне изображение не поменялось.

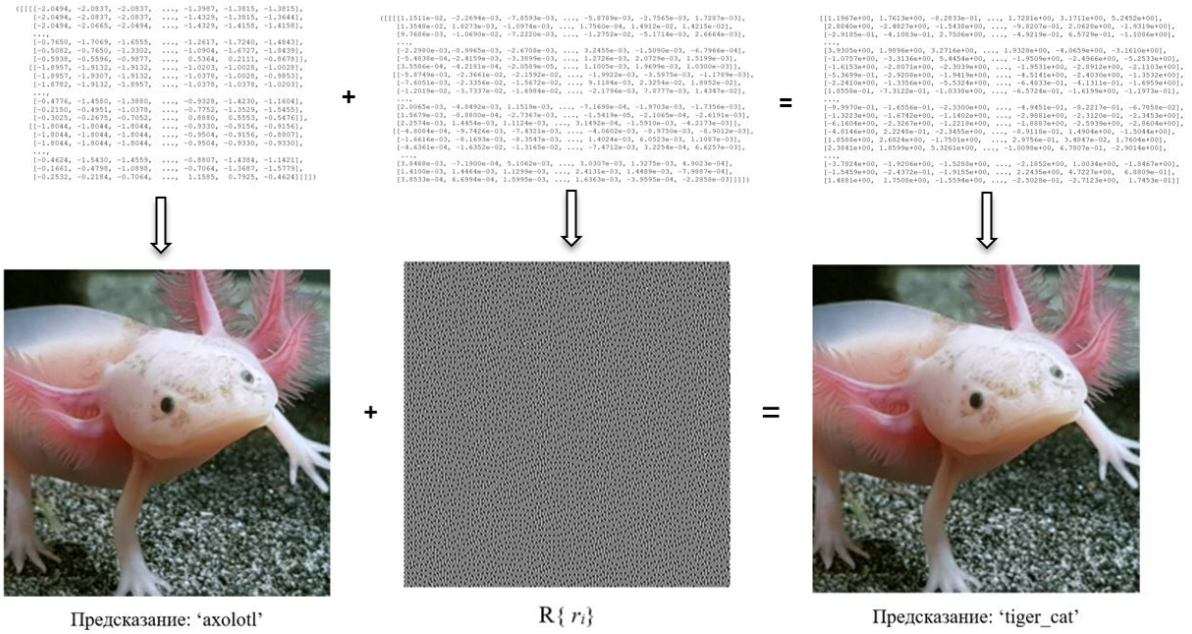


Рисунок 8 – Изображение с аксолотлем до и после искажения
Figure 8 – The image from the axolotl before and after doistortion

Однако нейронная сеть искаженное изображение распознает уже иначе (Рисунок 9).

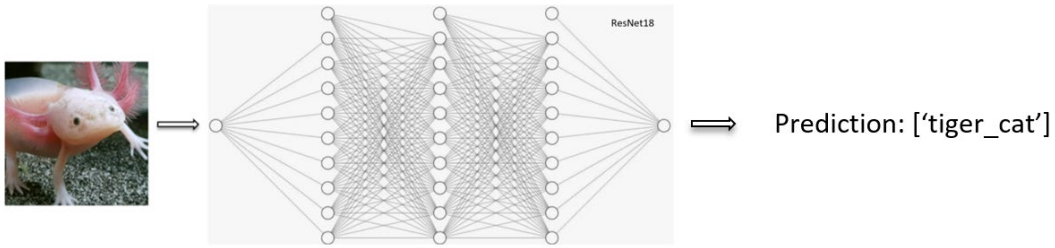


Рисунок 9 – Результат прогнозирования нейронной сетью модифицированного изображения
Figure 9 – The result of prediction by a neural network of a modified image

Аналогичным образом могут быть осуществлены атаки на другие информационные системы распознавания изображений, основанных на других нейронных сетях.

Матрица превращений

Для обучающего множества фиксированного размера, состоящего из изображений, может быть проведено исследование, направленное на изучение процесса взаимного «превращения» объектов. Полученная информация может быть сведена в матрицу превращений T, где $t_{i,j}$ – количество эпох минимально необходимых для «превращения» изображения с объектом из i-ого в j-ое. По диагонали матрицы количество эпох равно нулю, потому что каждый объект уже «превращен» сам в себя и дальнейшие трансформации не требуются.

Анализ данной матрицы может показать, какое «превращение» занимает минимальное / максимальное количество эпох, в какой объект в среднем в рамках данного обучающего множества проще / сложнее «превратить» любой другой объект, какие объекты в среднем в рамках данного обучающего множества «превращаются» друг в друга с одинаковой сложностью (при примерном равном количестве эпох).

Анализ результатов экспериментов

Был взят датасет из 11 животных и проведена серия экспериментов по модификации изображений таким образом, чтобы каждое животное распознавалось как каждое из оставшихся. Результат показан на диаграмме (Рисунок 10).

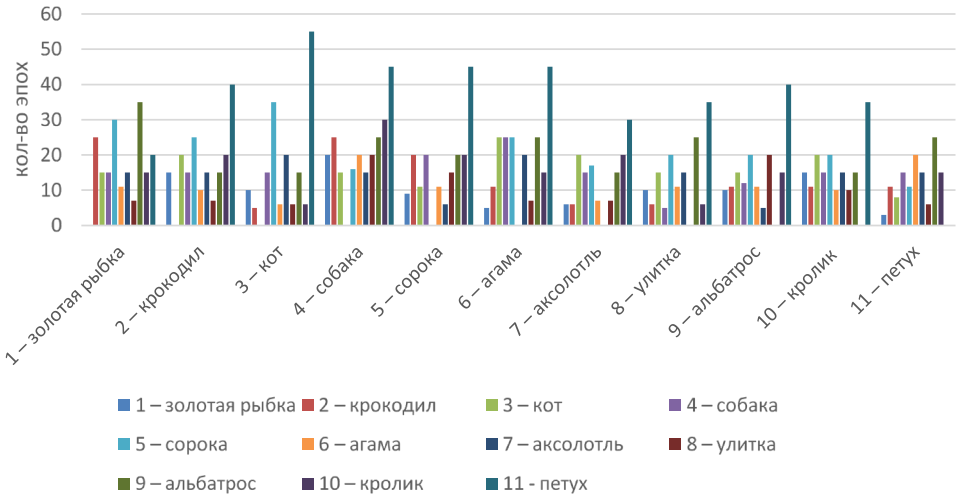


Рисунок 10 – Диаграммы преобразований
Figure 10 – Transformation diagrams

В Таблице 1 приведена матрица взаимных превращений T для данного обучающего множества. Количество эпох, которое необходимо для успешной атаки варьируется в диапазоне от 3 до 55.

Таблица 1 – Матрица «взаимных превращений»
Table 1 – Matrix of «mutual transformations»

	1	2	3	4	5	6	7	8	9	10	11
1	0	15	10	20	9	5	6	10	10	15	3
2	25	0	5	25	20	11	6	6	11	11	11
3	15	20	0	15	11	25	20	15	15	20	8
4	15	15	15	0	20	25	15	5	12	15	15
5	30	25	35	16	0	25	17	20	20	20	11
6	11	10	6	20	11	0	7	11	11	10	20
7	15	15	20	15	6	20	0	15	5	15	15
8	7	7	6	20	15	7	7	0	20	10	6
9	35	15	15	25	20	25	15	25	0	15	25
10	15	20	6	30	20	15	20	6	15	0	15
11	20	40	55	45	45	45	30	35	40	35	0

Анализируя матрицу T, можно выделить некоторые группы изображений:

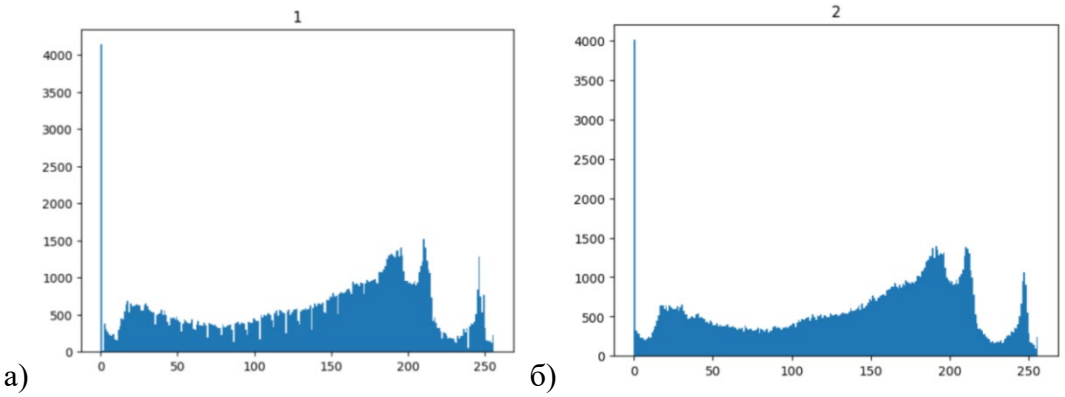
1) Группы, для взаимного превращения которых требуется примерно равное количество эпох. Например, чтобы преобразовать золотую рыбку (№ 1) в кролика (№ 10) потребуется 15 эпох, так же и из кролика в золотую рыбку.

2) Группы, для взаимного превращения которых требуется существенно разное количество эпох. Например, для преобразования петуха (№ 11) в кота (№ 3) требуется 55 эпох, а из кота в петуха – 8 эпох.

Анализ данных, приведенных в матрице, показывает, что для преобразования петуха в кошку требуется больше эпох, нежели преобразование агамы в кошку (Рисунок 11). Это связано с тем, что изображение с петухом имеет более насыщенную палитру цветов, в отличие от агамы. В связи с этим его сложнее изменить, так чтобы изменения были минимальными и не видимыми для человека.

Рисунок 11 – Изображения с разной цветовой палитрой
Figure 11 – Images with different color palettes

Сравним гистограмму трех цветовых каналов (RGB) петуха до совершения атаки (Рисунок 12 а) и после (Рисунок 12 б). По оси абсциссы гистограммы – значения от 0 до 255, которое равно значениям RGB каналов пикселя, по оси ординаты – значения от 0 до 4000, которое равно сумме всех каналов в данном значении. Для получения гистограммы изображение было разбито на RGB каналы, и была получена сумма количества всех каналов в каждом значении от 0 до 4000.

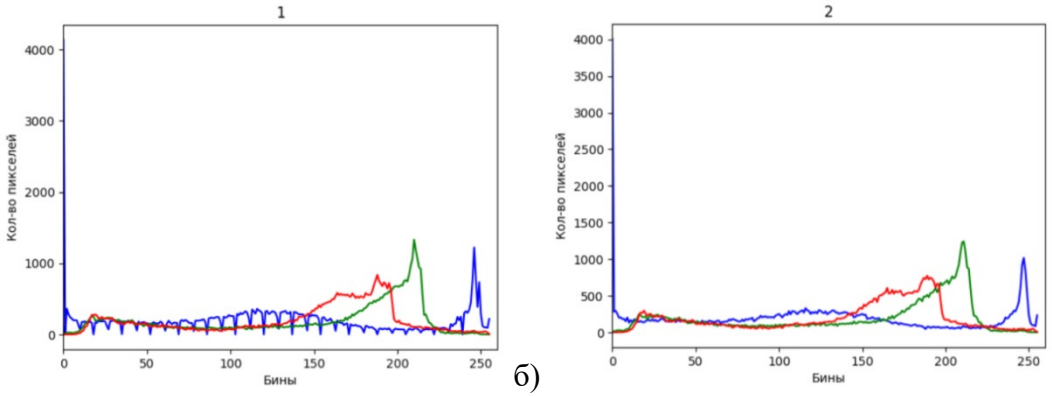


а) б)

Рисунок 12 – Гистограмма до (а) и после (б) атаки
Figure 12 – Histogram before (a) and after (b) the attack

Из приведенных гистограмм видно, что после атаки количество пикселей было изменено для некоторых значений, и теперь гистограмма изображения после атаки имеет более плавный переход между бинами.

Посмотрим на графики синего, зеленого и красного каналов (Рисунок 13).



а) б)

Рисунок 13 – Гистограмма 3 каналов до (а) и после (б) атаки
Figure 13 – Histogram of 3 channels before (a) and after (b) the attack

Из приведенных графиков видно, что количество пикселей было выровнено и теперь не имеет сильного разрыва в значениях. Наибольшие количество пикселей на ось ординат имеют те значения, которые отображают сам объект.

Заключение

Задачи распознавания изображений очень важны и применяются в самых разных областях. В рамках данной статьи были продемонстрированы примеры состязательных атак на сверточную нейронную сеть ResNet18. Было выявлено, что для успешной атаки может быть достаточно всего 3 итерации. Максимальное количество эпох в рамках данного исследования составило 55. Приведенный сравнительный анализ графиков трех

каналов цвета показал, как именно изменяется изображение в результате атаки. В дальнейшем планируется провести аналогичное исследование на системе распознавания лиц и разработать соответствующие механизмы защиты.

СПИСОК ИСТОЧНИКОВ / REFERENCES

1. Мурзина Д.О., Долженкова И.В. Применение систем искусственного интеллекта. *Форум молодых ученых.* 2017;(12):1313–1316.
   Murzina D.O., Dolzhenkova I.V. Application of artificial intelligence systems. *Forum molodykh uchenykh.* 2017;(12):1313–1316. (In Russ.).
2. Аликперова Н.В. Искусственный интеллект в здравоохранении: риски и возможности. *Здоровье мегаполиса.* 2023;4(3):41–49. https://doi.org/10.47619/2713-2617.zm.2023.v.4i3;41-49.
   Alikperova N.V. Artificial Intelligence in Healthcare: Risks and Opportunities. *Zdorov'e megapolisa = City Healtyhcare.* 2023;4(3):41–49. (In Russ.). https://doi.org/10.47619/2713-2617.zm.2023.v.4i3;41-49.
3. Прокопеня А.С., Азаров И.С. Сверточные нейронные сети для распознавания изображений. *BIG DATA and Advanced Analytics.* 2020;(6-1):271–280.
   Prokopenya A.S., Azarov I.S. Overview of convolutional neural networks for image recognition. *BIG DATA and Advanced Analytics*. 2020;(6-1):271–280. (In Russ.).
4. Назаров А.В., Марьенков А.Н., Калиев А.Б. Выявление поведенческих признаков работы вируса-шифровальщика на основе анализа изменений значений параметров компьютерной системы. *Прикаспийский журнал: управление и высокие технологии.* 2018;(1):196–204.
   Nazarov A.V., Marenkov A.N., Kaliev A.B. Detection of cryptographic viruses behavior signs in the work of the computer system. *Prikaspiiskii zhurnal: upravlenie i vysokie tekhnologii = Caspian Journal: Management and High Technologies.* 2018;(1):196–204. (In Russ.).
5. Марьенков А.Н., Кузнецова В.Ю., Гелагаев Т.М. Применение технологий распознавания лиц в системах контроля и управления доступом. *Прикаспийский журнал: управление и высокие технологии*. 2021;(1):83–90. https://doi.org/10.21672/2074-1707.2021.53.1.083-090.
   Marenkov A.N., Kuznetsova V.Yu., Gelagaev T.M. Application of face recognition technologies in control and access control systems. *Prikaspiiskii zhurnal: upravlenie i vysokie tekhnologii = Caspian Journal: Management and High Technologies.* 2021;(1):83–90. (In Russ.). https://doi.org/10.21672/2074-1707.2021.53.1.083-090.
6. Алексеенко Ю.В. Разработка системы распознавания изображений на основе сверточных нейронных сетей. *Евразийский Союз Ученых.* 2017;(7-1):8–11.
   Alekseenko Yu.V. Razrabotka sistemy raspoznavaniya izobrazhenii na osnove svertochnykh neironnykh setei. *Evraziiskii Soyuz Uchenykh.* 2017;(7-1):8–11. (In Russ.).
7. Демина Р.Ю., Ажмухамедов И.М. Повышение качества классификации объектов на основе введения новой метрики кластеризации. *Вестник Астраханского государственного технического университета. Серия: Управление, вычислительная техника и информатика.* 2019;(4):106–114. https://doi.org/10.24143/2072-9502-2019-4-106-114.
   Demina R.Yu., Azhmukhamedov I.M. Increasing quality of classifying objects using new metrics of clustering. *Vestnik Astrakhanskogo gosudarstvennogo tekhnicheskogo universiteta. Seriya: Upravlenie, vychislitel'naya tekhnika i informatika = Vestnik of Astrakhan State Technical University. Series: Management, Computer Science and*

Informatics. 2019;(4):106–114. (In Russ.). https://doi.org/10.24143/2072-9502-2019-4-106-114.

8. Чехонина Е.А., Костюмов В.В. Обзор состязательных атак и методов защиты для детекторов объектов. *International Journal of Open Information Technologies.* 2023;11(7):11–20.
   Chekhonina E.A., Kostyumov V.V. Overview of adversarial attacks and defenses for object detectors. *International Journal of Open Information Technologies.* 2023;11(7):11–20. (In Russ.).
9. Sai Abhishek A.V., Gurrala V.R., Sahoo L. Resnet18 Model With Sequential Layer For Computing Accuracy On Image Classification Dataset. *International Journal of Creative Research Thoughts.* 2022;10(5):176–181.
10. Сикорский О.С. Обзор свёрточных нейронных сетей для задачи классификации изображений. В сборнике: *Двадцатый научно-практический семинар «Новые информационные технологии в автоматизированных системах»: Материалы двадцатого научно-практического семинара, 20 апреля 2017 года, Москва, Россия.* Москва: Московский институт электроники и математики им. А.Н. Тихонова; 2017. С. 37–42.
    Sikorskii O.S. Obzor svertochnykh neironnykh setei dlya zadachi klassifikatsii izobrazhenii. In: *Dvadtsatyi nauchno-prakticheskii seminar «Novye informatsionnye tekhnologii v avtomatizirovannykh sistemakh»: Materialy dvadtsatogo nauchno-prakticheskogo seminara, 20 April 2017, Moscow, Russia.* Moscow: HSE Tikhonov Moscow Institute of Electronics and Mathematics; 2017. P. 37–42. (In Russ.).

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Хмелёва Анастасия Александровна, студентка, Факультет цифровых технологий и кибербезопасности, Астраханский государственный университет им. В.Н. Татищева, Астрахань, Российская Федерация.
e-mail: nastakhmeleva99@mail.ru
ORCID: 0009-0005-1465-1036

Anastasia A. Khmeleva, Undergraduate Student, Faculty of Digital Technologies and Cybersecurity, Astrakhan State University named after V.N. Tatishchev, Astrakhan, the Russian Federation.

Демина Раиса Юрьевна, кандидат технических наук, доцент Астраханского государственного университета им. В.Н. Татищева, Астрахань, Российская Федерация.
e-mail: raisa.demina.91@mail.ru
ORCID: 0009-0009-1615-5641

Raisa Y. Demina, Candidate of Engineering Sciences, Associate Professor of Astrakhan State University named after V.N. Tatishchev, Astrakhan, the Russian Federation.

Ажмухамедов Искандар Маратович, доктор технических наук, профессор Астраханского государственного университета им. В.Н. Татищева, Астрахань, Российская Федерация.
e-mail: aim_agtu@mail.ru
ORCID: 0000-0001-9058-123X

Iskandar M. Azhmukhamedov, Doctor of Engineering Sciences, Professor at Tatishchev Astrakhan State University named after V.N. Tatishchev, Astrakhan, the Russian Federation.

Статья поступила в редакцию 04.04.2024; одобрена после рецензирования 15.04.2024; принята к публикации 19.04.2024.

The article was submitted 04.042024; approved after reviewing 15.04.2024; accepted for publication 19.04.2024.