ТЕХНИЧЕСКИЕ НАУКИ

TECHNICAL SCIENCES

Научная статья
УДК 004.032.26.056+519.7
ББК 32.972.13+22.18
Ч 48
DOI: 10.53598 / 2410-3225-2021-4-291-52-57

Разработка системы обеспечения безопасности распознавания образов от вредоносного воздействия на основе нейронной сети

(*Рецензирована*)

Александр Николаевич Черкасов[1], Евгений Евгеньевич Волнянкин[2]

[1, 2] *Кубанский государственный технологический университет, Краснодар, Россия*
[1] *cherk@mail.ru*
[2] *cyber21fox@gmail.com*

Аннотация. *Представлены результаты исследования вредоносного воздействия на изображения и предложен алгоритм защиты для систем компьютерного зрения. Предложена интеллектуальная система защиты и распознавания образов на основе двух типов нейронных сетей: автоэнкодера и сверточной нейронной сети.*

Ключевые слова: *сверточная нейронная сеть, автоэнкодер, распознавание образов, гауссовское зашумление, создание устойчивой модели*

Original Research Paper

Development of a system for ensuring the safety of image recognition from malicious influences based on a neural network

Aleksandr N. Cherkasov[1], Evgeniy E. Volnyankin[2]

[1, 2] *Kuban State University of Technology, Krasnodar, Russia*
[1] *cherk@mail.ru*
[2] *cyber21fox@gmail.com*

Abstract. *The paper presents the results of a study of the harmful effects on images and proposes a protection algorithm for computer vision systems. An intelligent protection and pattern recognition system, based on two types of neural networks: an autoencoder and a convolutional neural network, is proposed.*

Keywords: *convolutional neural network, autoencoder, pattern recognition, Gaussian noise, stable model creation*

Использование новых технологий, связанных с искусственным интеллектом и машинным обучением, активно применяется во многих областях деятельности человека [1]. Новые алгоритмы помогают распознавать знаки дорожного движения, водить беспилотники, фильтровать спам, распознавать лица и т.д. Однако с развитием таких технологий встает вопрос, а насколько эти технологии безопасны и насколько им можно доверять. По мере роста потребления продуктов и услуг, созданных на основе машинного обучения, необходимо предпринимать специальные меры, чтобы защитить не

только пользователей и их данные, но и сам искусственный интеллект и алгоритмы от нарушения работоспособности. Поэтому в данный момент актуальной становится задача обеспечения безопасности алгоритмов машинного обучения и в частности систем компьютерного распознавания изображений [2].

Существует целая группа атак, которая вводит в заблуждение системы искусственного интеллекта. Самые известные виды атак, использующих вредоносный образ: FGSM, L-BFGS, DeepFool, JSMA, Onepixel [3]. Основная суть таких угроз заключается в возможности раскрытия нарушителем информации о модели машинного обучения, хищения нарушителем обучающих данных, модификации (искажения) модели машинного обучения. На рисунке 1 приведена структура проведения атаки на систему машинного обучения, в том числе на системы распознавания образов, детекции объектов и т.д. [4].

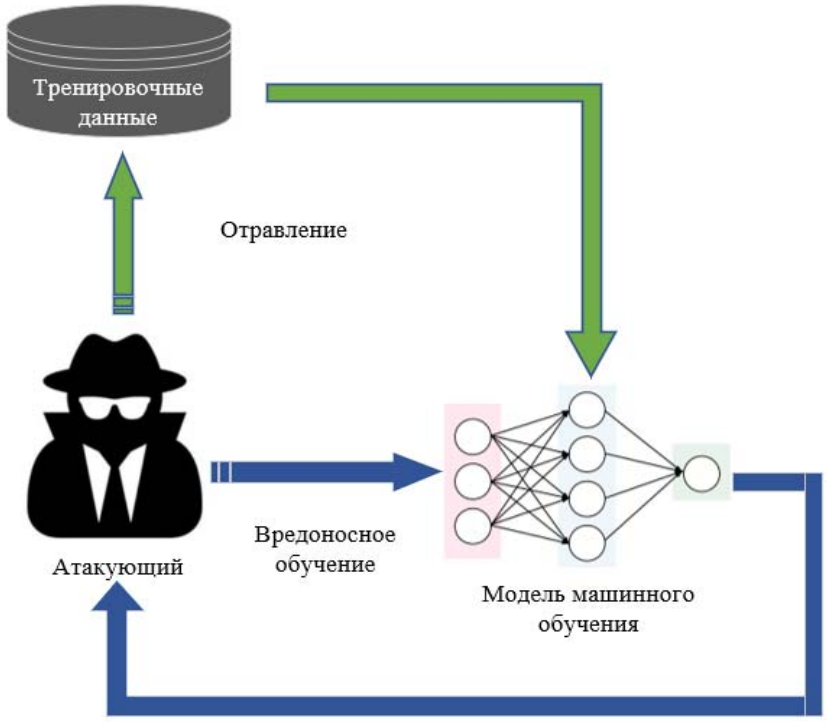

Рис. 1. Схема атаки на модель машинного обучения
Fig. 1. Pattern of attack on machine learning model

Тема уязвимостей систем машинного обучения, искусственных нейронных сетей к атакам в настоящее момент недостаточно изучена, а универсальных методов защиты, способных решить вышеуказанные проблемы, пока нет [5]. Однако разработаны некоторые подходы, позволяющие сгладить последствия атак на искусственные нейронные сети (ИНС) [6].

В статье приведена модель улучшения по распознаванию атакованного изображения, функционирующая на основе устойчивой искусственной нейронной сети. Авторами разработана модель, позволяющая на основе генерации вредоносных образов, сформированных на основе гауссовского шума, распознавать исходное изображение. Генерация вредоносных образов на набор данных и дальнейшее обучение позволяет увеличить устойчивость модели.

Схема работы алгоритма приведена на рисунке 2.

С целью реализации алгоритма разработана программа, представляющая собой схему двух моделей, построенных на двух типах нейронных сетей: автоэнкодер и сверточная нейронная сеть.

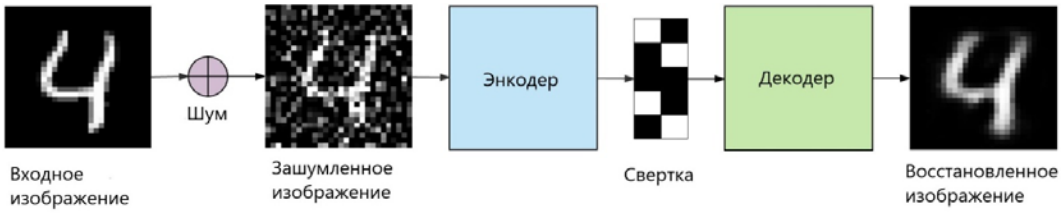

Рис. 2. Алгоритм улучшения модели устойчивости нейронной сети
Fig. 2. Algorithm for improving neural network resistance model

Первая модель – сверточная нейронная сеть, которая в качестве классификатора исполняет роль системы распознавания образов [7]. Второй моделью, которая исполняет роль системы защиты, является автоэнкодер [8]. Он исполняет как и основную защитную роль сжатия и восстановления изображения, так и другие. На рисунке 3 приведена схема слоев сверточной нейронной сети.

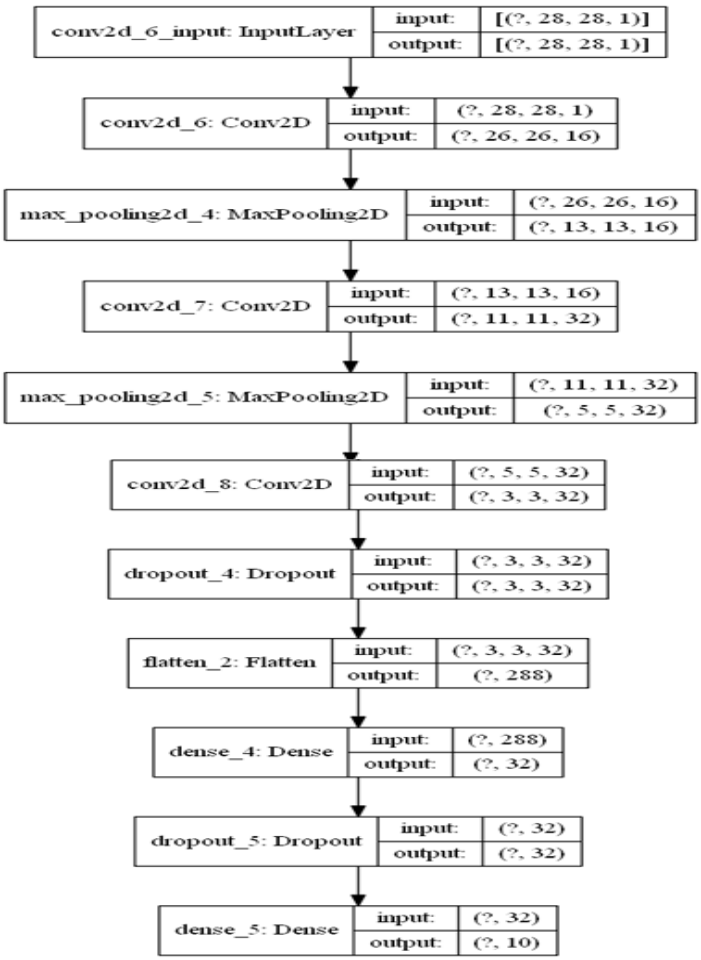

Рис. 3. Схема слоев сверточной нейронной сети
Fig. 3. Diagram of convolutional neural network layers

На рисунке 4 приведена схема нейронной сети автоэнкодера, представляющего собой энкодер, скрытое пространство и декодер.

Для проверки функционирования программы использовалась одна из наиболее распространенных атак: метод быстрого градиентного знака, работающий с использованием градиентов нейронной сети для создания состязательного примера.

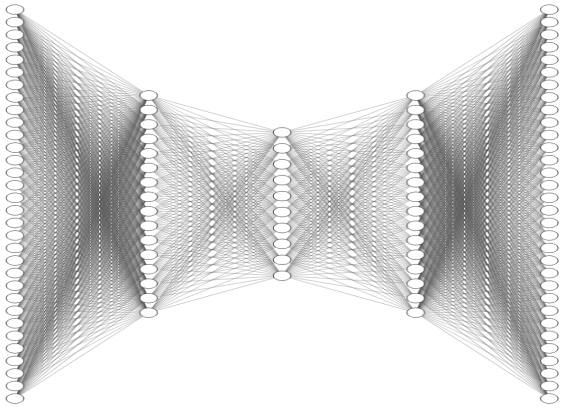

Рис. 4. Схема слоев автоэнкодера
Fig. 4. Autoencoder layer diagram

Для входного изображения метод использует градиенты потерь по отношению к входному изображению, чтобы создать новое изображение, которое максимизирует потери. Для этого нужно определить, насколько каждый пиксель изображения влияет на величину потерь. Атаку можно описать формулой:

$$\text{adv_x} = \text{x} + \text{e} \cdot \text{sing}(\text{J}(\text{и}, \text{x}, \text{y})),$$

где adv_x – вредоносное изображение;
x – исходное изображение;
e – множитель для обеспечения зашумления;
J – потеря;
и – параметры модели;
y – метка изображения [9].

Для тестирования разработанного алгоритма и программы была выбрана база данных рукописных цифр MNIST (odified National Institute of Standards and Technology). Данные состоят из заранее подготовленных примеров изображений, на основе которых проводится обучение и тестирование систем. Ее преимуществом является то, что набор данных занимает небольшое количество памяти и изображения малого размера. Набор состоит из 60000 полутоновых изображений размером 28×28. Всего в наборе 10 цифр. Тестовый набор состоит из 10000 изображений. Фрагмент представлен на рисунке 5.

Рис. 5. Фрагмент набора исходных данных изображения
Fig. 5. Snippet of image source data

Сверточная нейронная сеть после прохождения через нее изображения в результате выдает метку класса изображения: изображению присваиваются значения от 0 до 9 (см. рис. 6).

Рис. 6. Тестовые изображение с метками классов
Fig. 6. Test image with class labels

На созданный набор данных на основе тестовых значений и зашумленных данных была проведена атака по методу быстрого градиентного знака (FGSM) с помощью библиотеки ART-IBM.

Пример реализации атаки FGSM при эпсилоне, равном 0.001, приведен на рисунке 7.

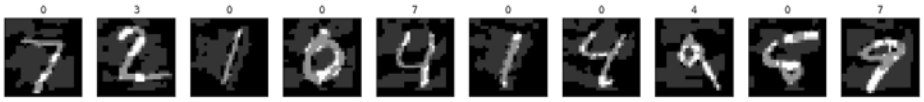

Рис. 7. Пример изображения с наложением, основанном на методе градиентного знака
Fig. 7. Example of image with overlay based on gradient sign method

После успешно проведенной атаки заметно изменяются изображения, а с ними и метки классов.

Применяя разработанную программу, состоящую из сверточной нейронной сети и автоэнкодера с добавление зашумленных объектов в обучающую выборку, был получен следующий результат, приведенный на рисунке 8.

Рис. 8. Результат работы разработанной программы
Fig. 8. Result of the developed program operation

Функционирование разработанной системы представляет собой наглядный пример того, что при введении в любую систему компьютерного зрения, обладающую своими свойствами, фильтра в виде автоэнкодера, появляется возможность обеспечить дополнительную защиту и снизить вероятность успешной атаки на систему машинного обучения. После нормализации данных, составления модели и обучения точность классификации сверточной нейронной сетью составила 98,03%.

Выводы:

1. Предлагаемый алгоритм распознавания изображений на основе функционирования двух видов нейронных сетей обеспечивает достаточную степень защиты от атак и вредоносного обучения. Стоит отметить, что данная схема может с успехом применяться для решения более широкого круга задач.

2. Разработанная интеллектуальная система по улучшению атакованного изображения на основе представленных выше результатов позволяет защититься от аномальных входных данных с большим зашумлением, способных «отравить» модель.

Примечания

1. Черкасов А.Н. Разработка математического и алгоритмического обеспечения адаптивных систем поддержки принятия решений в ситуационных центрах: дис. … канд. техн. наук. Краснодар: Кубанский государственный технологический университет, 2011. 152 с.

2. Кэти Уорр. Надежность нейронных сетей. Укрепляем устойчивость ИИ к обману. Санкт-Петербург: Питер, 2021. 272 с.

3. Безопасность алгоритмов машинного обучения. Атаки с использованием Python.

URL: https://dsec.ru/research-and-analytics/article/bezopasnost-algoritmov-mashinnogo-obucheniya-zashhita-i-testirovanie-modelej-s-ispolzovaniem-python/ (дата обращения: 03.06.2021).

4. Комплексная платформа машинного обучения с открытым исходным кодом. URL: https://www.tensorflow.org/ (дата обращения: 03.06.2021).

5. Черкасов А.Н., Туркин Е.А. Разработка модели обнаружения вредоносных программ на основе анализа последовательностей API-запросов // Вестник Адыгейского государственного университета. Сер.: Естественно-математические и технические науки. 2021. Вып. 2 (281). С. 90–96. URL: http://vestnik.adygnet.ru

6. Кларенс Чио, Дэвид Фримэн. Машинное обучение и безопасность. Москва: ДМК Пресс, 2020. 388 с.

7. Черкасов А.Н., Сивенко А.В. Разработка модели обеспечения безопасности аккаунта социальной сети на основе нейросетевого алгоритма // Вестник Адыгейского государственного университета. Сер.: Естественно-математические и технические науки. 2021. Вып. 2 (281). С. 47–55. URL: http://vestnik.adygnet.ru

8. Николенко С., Кадурин А., Архангельская Е. Глубокое обучение. Погружение в мир нейронных сетей. Москва, 2018. 477 с.

9. Плас Дж. Вандер. Python для сложных задач: наука о данных и машинное обучение. Санкт-Петербург: Питер, 2018. 576 с.

References

1. Cherkasov A.N. Development of mathematical and algorithmic support for adaptive decision support systems in situational centers: Diss. for the Cand. of Techn. Sciences degree. Krasnodar: Kuban State University of Technology, 2011. 152 p.

2. Katy Warr. Strengthening Deep Neural Networks: Making AI Less Susceptible to Adversarial Trickery. St. Petersburg: Piter, 2021. 272 p.

3. Security of machine learning algorithms. Python Attacks. URL: https://dsec.ru/research-and-analytics/article/bezopasnost-algoritmov-mashinnogo-obucheniya-zashhita-i-testirovanie-modelej-s-ispolzovaniem-python/ (access date: 03.06.2021).

4. A comprehensive open source machine learning platform. URL: https://www.tensorflow.org/ (access date: 03.06.2021).

5. Cherkasov A.N., Turkin E.A. Development of the malware detection model based on an analysis of API request sequences // The Bulletin of the Adyghe State University. Ser.: Natural-Mathematical and Technical Sciences. 2021. Iss. 2 (281). P. 90–96. URL: http://vestnik.adygnet.ru

6. Clarence Chio, David Freeman. Machine learning and security. Moscow: DMK Press, 2020. 388 p.

7. Cherkasov A.N., Sivenko A.V. Development of a model for ensuring the security of a social network account based on a neural network algorithm // The Bulletin of the Adyghe State University. Ser.: Natural-Mathematical and Technical Sciences. 2021. Iss. 2 (281). P. 47–55. URL: http://vestnik.adygnet.ru

8. Nikolenko S., Kadurin A., Arkhangelskaya E. Deep Learning. Dive into the world of neural networks. Moscow, 2018. 477 p.

9. Plas J. Wonder. Python for Complex Problems: Data Science and Machine Learning. St. Petersburg: Piter, 2018. 576 p.

Авторы заявляют об отсутствии конфликта интересов.

The authors declare no conflicts of interests.

Статья поступила в редакцию 11.11.2021; одобрена после рецензирования 2.12.2021; принята к публикации 3.12.2021.

The article was submitted 11.11.2021; approved after reviewing 2.12.2021; accepted for publication 3.12.2021.

© А.Н. Черкасов, Е.Е. Волнянкин, 2021