DOI 10.37882/2223-2966.2023.7-2.34

АЛГОРИТМ ЗАЩИТЫ СИСТЕМ МАШИННОГО ОБУЧЕНИЯ ОТ УГРОЗЫ МОДИФИКАЦИИ МОДЕЛИ ПУТЕМ ВЫЯВЛЕНИЯ ВНЕДРЕННЫХ ВРЕДОНОСНЫХ ДАННЫХ

Чекмарев Максим Алексеевич
Адъюнкт, Краснодарское высшее военное училище им. С.М. Штеменко
max.chek13@gmail.com

Аннотация: В работе рассмотрены основные угрозы безопасности для систем машинного обучения и, в частности, угрозы модификации модели путем внедрения вредоносных данных. Предложен алгоритм защиты, основанный на применении технологии развертывающихся искусственных нейронных сетей. Проведено тестирование работы алгоритма с использованием разработанной программы для ЭВМ и получен набор низкоуровневых признаков объекта обучения, позволяющий выявить внедренные вредоносные данные. Сделаны выводы о необходимости проведения дальнейших исследований по данному направлению.

Ключевые слова: машинное обучение, искусственный интеллект, нарушение безопасности, атака отравления, искусственные нейронные сети, низкоуровневые признаки.

AN ALGORITHM TO PROTECT MACHINE LEARNING SYSTEMS FROM THE THREAT OF MODEL MODIFICATION BY DETECTING EMBEDDED MALICIOUS DATA

M. Chekmarev

Summary. The paper discusses the main security threats to machine learning systems and, in particular, the threat of model modification through the introduction of malicious data. A security algorithm based on the deployment technology of artificial neural networks is proposed. The algorithm has been tested using the developed computer program and a set of low-level features of the learning object, which allows to identify the embedded malicious data, has been obtained. Conclusions are drawn on the need for further research in this area.

Keywords: machine learning, artificial intelligence, security breanch, poisoning attack, artificial neural networks, low-level features.

Уязвимости систем машинного обучения

Важность развития искусственного интеллекта в целях обеспечения национальных интересов и реализации стратегических национальных приоритетов (в том числе в области научно-технологического развития) Российской Федерации определена Национальной стратегией развития искусственного интеллекта на период до 2030 года [1]. Необходимость широкого внедрения и применения систем искусственного интеллекта и машинного обучения отмечена в решениях Президента Российской Федерации в ходе проведения отдельных выступлений [2].

Вместе с тем, применение систем машинного обучения приносит с собой проблемы обеспечения их защиты от атак манипуляции данными на всех этапах жизненного цикла, которые способные привести к нарушениям конфиденциальности, целостности и доступности информации, что в конечном итоге может привести к неправильному функционированию технических средств и систем, в которых они применяются.

Таксономия атак на системы машинного обучения подробно описана в [3], дополнена в [4]. Кроме этого, сценарии угроз информационной безопасности для таких систем внесены в Банк данных угроз безопасности информации ФСТЭК России под индексами УБИ.218-УБИ.222 [5].

В общем виде атаки на системы машинного обучения можно разделить на атаки, преследующие целью воздействие на обучающие данные (на этапе обучения модели), состязательные (в ходе эксплуатации обученной модели) и исследовательские атаки, при которых задача злоумышленника состоит в извлечении из модели (обучающих данных) конфиденциальной или другой, представляющей ценность, информации.

Учитывая, что обучение модели является первым этапом процесса жизненного цикла системы машинного обучения, решение задачи обеспечения безопасности в ходе него является критически важным вопросом.

Атака на набор обучающих данных может быть деструктивной и целевой. В первом случае злоумышленник в ходе формирования обучающих данных вводит метки для данных, не соответствующие этим данным, что приводит к модификации модели, изменению показателей метрик и неправильному ее поведению на реальных данных [6]. Целевая атака предполагает такое изменение совокупности признаков объектов обучения, незаметное при восприятии человеком, при котором обученная в дальнейшем модель машинного обучения будет верно

классифицировать входной объект, подверженный заражению. Однако, при этом относить к тому же классу и другие объекты, в которые внедрены аналогичные вредоносные данные.

Таким образом, если x — исходные входные данные, а r — вектор, представляющий собой изменение исходных входных данных, то вредоносные данные будет представлять собой сумму этих показателей $x_{врд} = x + r$. При этом результат выполнения функции, применяемой алгоритмом машинного обучения, от такого аргумента должен быть равен целевой категории вредоносной атаки $f(x_{врд}) = y_ц$.

Кроме этого, задача злоумышленника при целевой атаке сводится к минимизации величины вредоносного изменения r, чтобы сделать его незаметным для человека: $\operatorname{argmin}\left\{\|r\| : f\left(x_{врд}\right) = y_ц\right\}$.

Опасность целевой атаки заключается в том, что, если последствия деструктивной атаки проявляются достаточно заметно в ходе тестирования и эксплуатации в виде падения показателей метрик или выдаче неверных результатов, то в случае внедрения вредоносных данных такого не происходит. Это дает возможность злоумышленнику манипулировать моделью машинного обучения по своему умыслу.

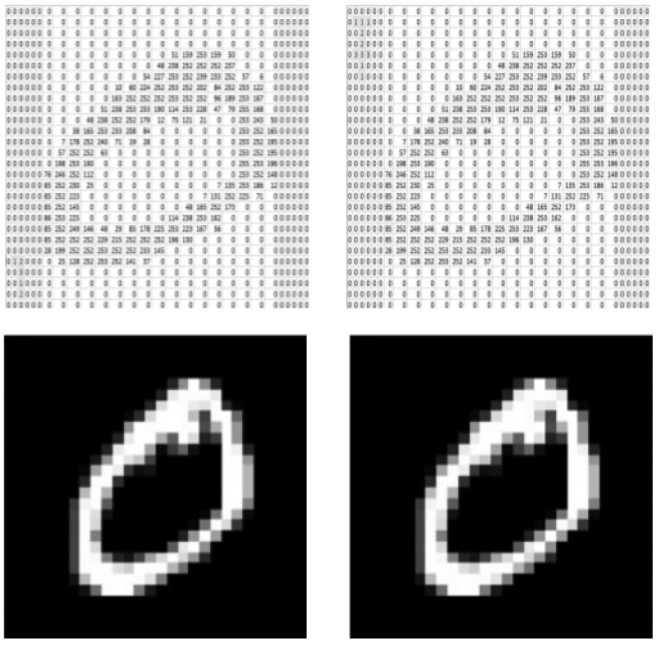

Рис. 1. Представление объекта «0»
с внедренными вредоносными данными в базе данных (сверху) и в виде графического отображения (снизу)

Так, для выявления последствий возможной атаки была создана база на основе набора рукописных цифр MNIST [7], состоящая из 5-ти наборов данных, в каждый из которых в вектор признаков объекта «0» внесены изменения в виде повторяющихся цифр, формирующих пиксели в оттенках серого (1, 2, 3, 5 и 10 пикселей соответственно), расположенных в определенной последовательности. При визуализации объекта данные изменения для человеческого глаза невосприимчивы (рис. 1). На каждом из наборов данных обучены 5 искусственных нейронных сетей.

Была создана тестовая выборка, в которой в каждый объект с меткой «7» внедрены те же данные, что и в объект «0» и осуществлено измерение точности распознавания каждой из моделей. При этом вероятность распознавания объекта «7» как объекта «0» при внедрении на обученной модели с внедренными десятью пикселями увеличивается в 30 тысяч раз.

Алгоритм

Для решения задачи выявления внедренных вредоносных данных разработан алгоритм, в основе которого лежат технологии глубокого обучения, а именно применение концепции и технологии развертывающихся нейронных сетей (Deconvolution Neural Networks), способные формировать структуры, позволяющие делать выводы о том, какие низкоуровневые признаки объекта являются ключевыми для его классификации.

Развертывающаяся нейронная сеть строит иерархические представления сверточной нейронной сети. Она подключается к каждому сверточному слою нейронной сети и восстанавливает изображения для всех сверточных слоев, обучаясь параллельно. В итоге получается нейронная сеть, которая позволяет «видеть» то, как обучена сверточная нейронная сеть и интерпретировать результаты [8].

В цикле анализа объектов, вложенном в цикл анализа групп объектов, последовательно проводятся операции, характерные для сверточных и развертывающихся нейронных сетей: свертка, вычисление карты признаков, обнуление весов нейронов, разъединение, ректификация и фильтрация. Результатом работы алгоритма является набор низкоуровневых признаков объекта, позволяющий выявить внедренные вредоносные данные (рис. 2).

С целью проверки работы алгоритма разработана программа, предназначенная для выявления вредоносных данных, внедренных в цифровые изображения, формируемые из файлов формата «csv» с возможностью выбора набора обучающих данных и количества эпох для обучения, обучения модели искусственной нейронной сети, формирования градиентных признаков изображения, вывода исходного изображения и изображения его низкоуровневых признаков в соответствии с заданной пользователем меткой.

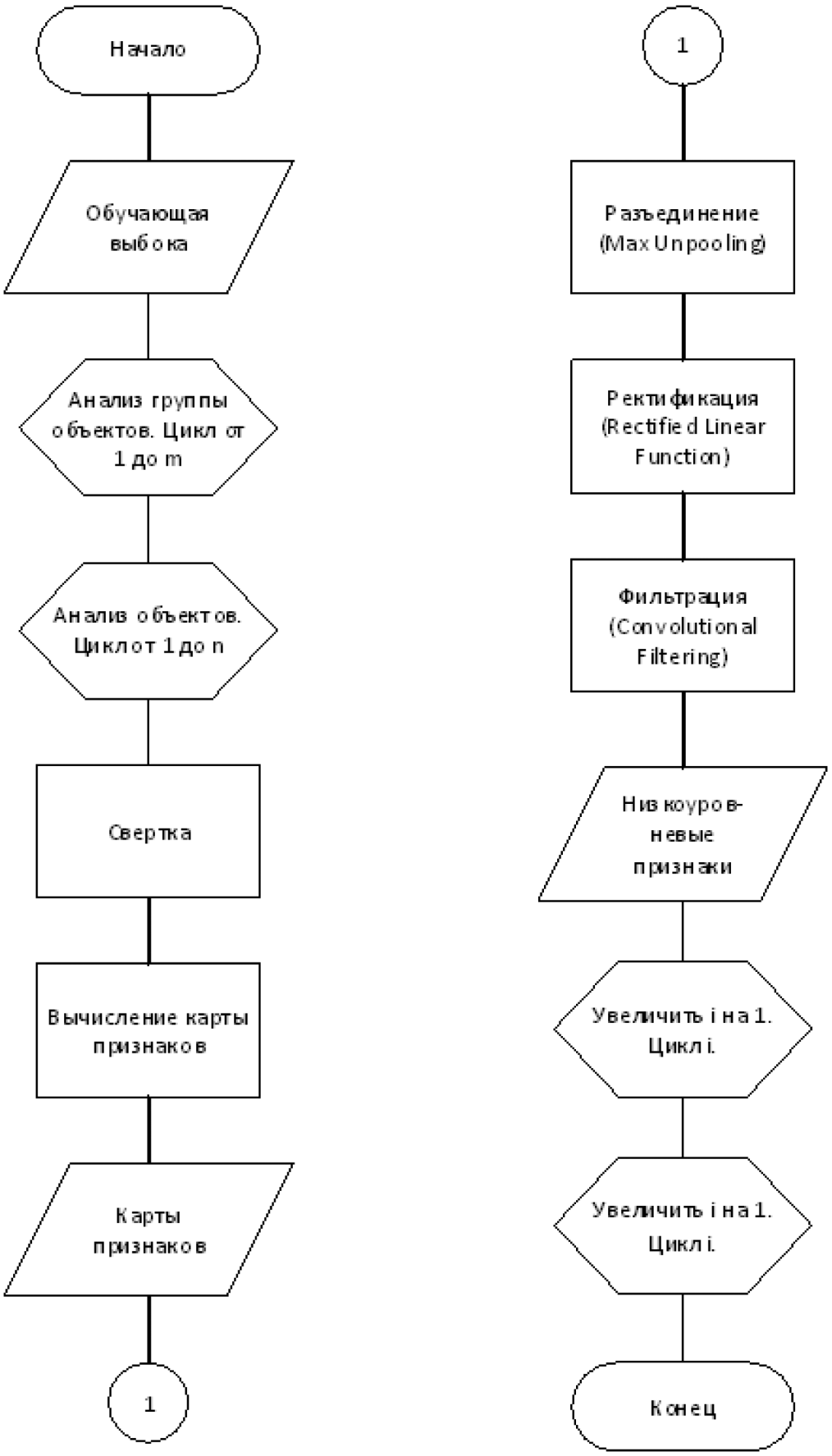

Рис. 2. Блок-схема работы алгоритма

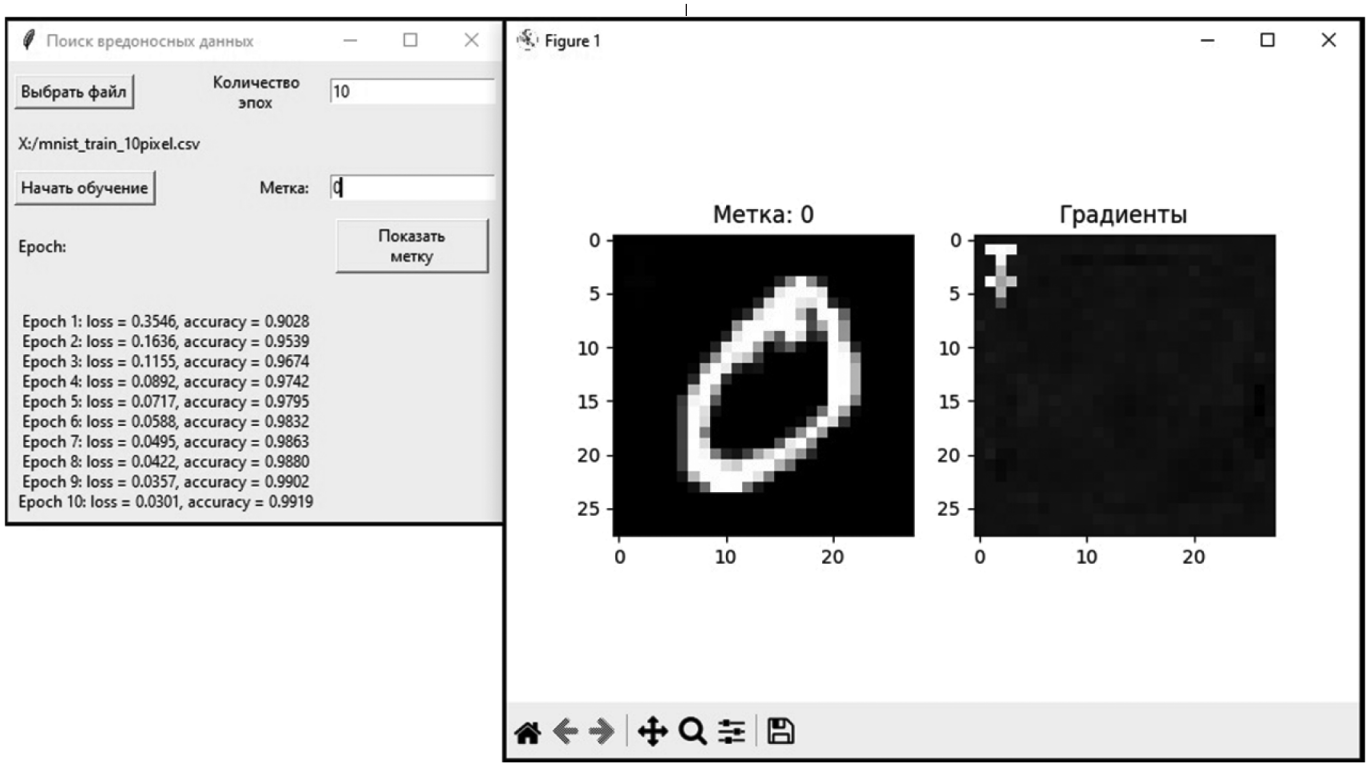

Рис. 3. Выявление внедренных данных (10 пикселей) в объект с меткой «0»

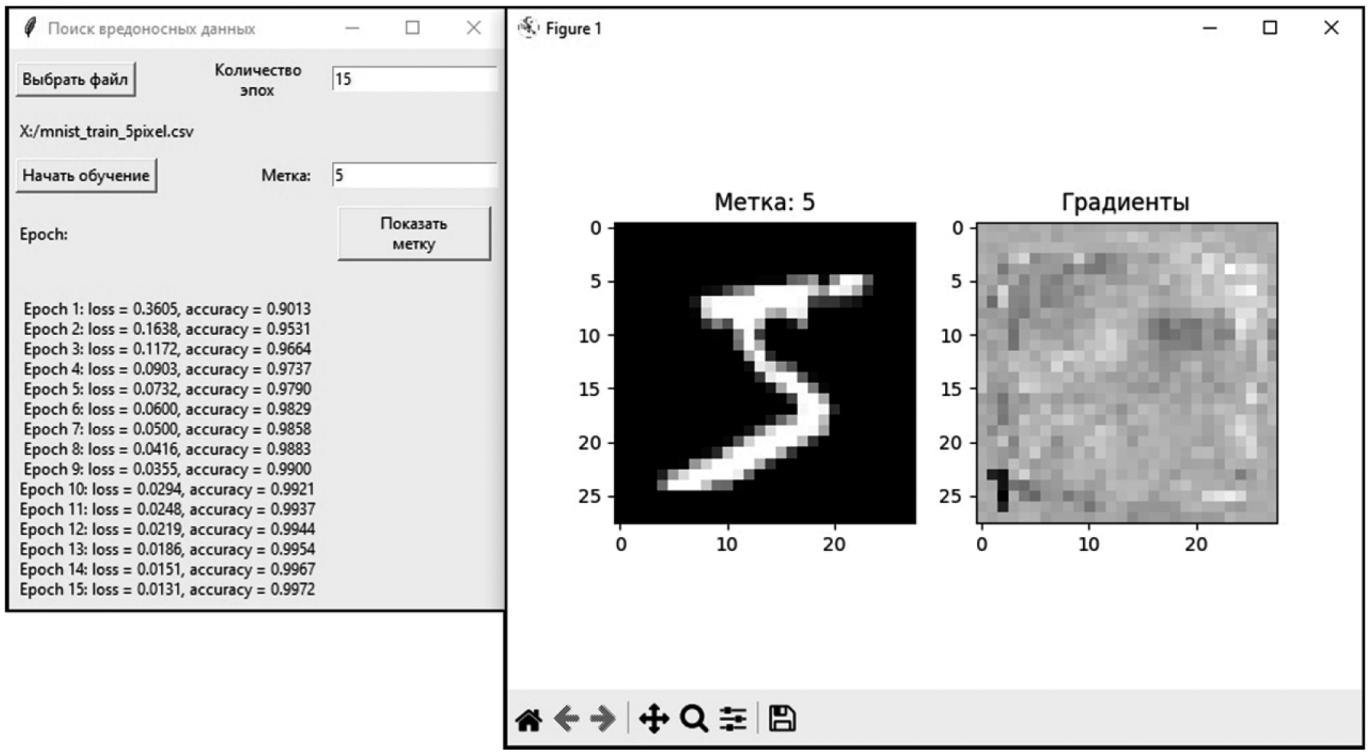

Рис. 4. Выявление внедренных данных (5 пикселей) в объект с меткой «5»

В ходе тестирования работы программы осуществлялось обучение двухслойной нейронной сети прямого распространения на наборе данных с внедренными вредоносными данными. Последующая визуализация низкоуровневых признаков объектов показало их явное наличие в том месте изображения, куда они были внедрены (рис. 3, 4).

Заключение

Задача защиты систем машинного обучения от угрозы модификации модели путем выявления внедренных вредоносных данных не сводится исключительно к выявлению таковых на цифровых изображениях. Она актуальна для систем интеллектуального анализа текста, аудио- и видеозаписей, классификации вредоносных файлов и т. д. Таким образом, проведение дальнейших исследований в данном направлении является актуальной задачей.

ЛИТЕРАТУРА

1. О развитии искусственного интеллекта в Российской Федерации: указ Президента РФ от 10.10.2019 г. № 490 // СПС «КонсультантПлюс». –Режим доступа: http://www.consultant.ru.
2. Путин призвал массово внедрить в этом десятилетии искусственный интеллект во все отрасли // ТАСС URL: https://tass.ru/ekonomika/16418761 (дата обращения: 04.05.2023).
3. A Taxonomy and Terminology of Adversarial Machine Learning. Draft NISTIR 8269, October 2019. Режим доступа: https://doi.org/10.6028/ NIST.IR.8269-draft.
4. Безопасность систем машинного обучения. защищаемые активы, уязвимости, модель нарушителя и угроз, таксономия атак / В.Г. Грибунин, Р.Л. Гришаненко, А.П. Лабазников, А.А. Тимонов // Известия Института инженерной физики. — 2021. — № 3(61). — С. 65–71.
5. Федеральная служба по таможенному и экспортному контролю (ФСТЭК России). Банк данных угроз безопасности информации [Электронный ресурс]. Режим доступа: http://bdu.fstec.ru/threat/, свободный. Яз. рус. (дата обращения: 04.05.2023).
6. Чекмарев, М.А. Экспериментальная проверка угрозы модификации модели машинного обучения в результате искажения обучающих данных / М.А. Чекмарев, Н.Д. Бобров // Информационно-телекоммуникационные системы и технологии: Материалы Всероссийской научно-практической конференции, Кемерово, 26 ноября 2021 года / Редколлегия: А.Г. Пимонов (отв. ред.) [и др.]. — Кемерово: Кузбасский государственный технический университет имени Т.Ф. Горбачева, 2021. — С. 31–34.
7. Свидетельство о государственной регистрации базы данных № 2022620255 Российская Федерация. Искаженная база обучающих данных для тестирования систем машинного обучения на предмет устойчивости к модификации модели: № 2021623139: заявл. 16.12.2021: опубл. 31.01.2022 / М. А. Чекмарев.
8. Душкин Р.В. Искусственный интеллект. — М.: ДМК Пресс, 2019. — 280 с.

Журнал «Современная наука: актуальные проблемы теории и практики»