УНИВЕРСИТЕТ ИТМО
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
июль–август 2022 Том 22 № 4 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
Jule–August 2022 Vol. 22 No 4 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК УНИВЕРСИТЕТ ИТМО
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2022-22-4-742-750
УДК 004.89

Метод защиты нейронных сетей от компьютерных бэкдор-атак на основе идентификации триггеров закладок

Артем Бакытжанович Менисов[1]✉, Александр Григорьевич Ломако[2], Андрей Сергеевич Дудкин[3]

[1,2,3] Военно-космическая академия имени А.Ф.Можайского, Санкт-Петербург, 197198, Российская Федерация

[1] vka@mil.ru✉, https://orcid.org/0000-0002-9955-2694
[2] vka@mil.ru, https://orcid.org/0000-0002-1764-1942
[3] vka@mil.ru, https://orcid.org/0000-0003-0283-9048

Аннотация

Предмет исследования. Современные технологии разработки и эксплуатации нейронных сетей уязвимы для компьютерных атак с внедрением программных закладок (бэкдор). Программные закладки могут оставаться скрытыми неопределенное время, пока не будут активированы вводом модифицированных данных, содержащих триггеры. Такие закладки представляют непосредственную угрозу безопасности информации для всех компонентов системы искусственного интеллекта. Такие воздействия злоумышленников приводят к ухудшению качества или полному прекращению функционирования систем искусственного интеллекта. В работе предложен оригинальный метод защиты нейронных сетей, сущность которого состоит в создании базы ранжированных синтезированных триггеров закладок целевого класса бэкдор-атак. **Метод.** Предложенный метод защиты нейронных сетей реализован путем последовательности защитных действий: выявлении закладки, идентификации триггера и нейтрализации закладки. **Основные результаты.** На основе представленного метода разработано программно-алгоритмическое обеспечение испытаний нейронных сетей, позволяющее выявить и нейтрализовать закладки для осуществления компьютерных бэкдор-атак. Экспериментальные исследования проведены на различных архитектурах сверточных нейронных сетей, обученных на наборах данных, для таких объектов, как аэрофотоснимки (DOTA), рукописные цифры (MNIST) и фотографии лиц людей (LFW). Снижение эффективности бэкдор-атак (не более 3 %) и малые потери качества функционирования нейронных сетей (на 8–10 % от качества функционирования нейронной сети без закладки) показало успешность разработанного метода. **Практическая значимость.** Применение предложенного метода защиты нейронных сетей позволит специалистам по информационной безопасности целенаправленно противодействовать компьютерным бэкдор-атакам на системы искусственного интеллекта и создать новые автоматизированные средства защиты информации.

Ключевые слова

искусственный интеллект, искусственная нейронная сеть, информационная безопасность, компьютерные атаки, бэкдор, закладки в нейронных сетях, синтезированные триггеры

Благодарности

Работа выполнена в рамках гранта Президента Российской Федерации для государственной поддержки молодых российских ученых — кандидатов наук МК-2485.2022.4.

Ссылка для цитирования: Менисов А.Б., Ломако А.Г., Дудкин А.С. Метод защиты нейронных сетей от компьютерных бэкдор-атак на основе идентификации триггеров закладок // Научно-технический вестник информационных технологий, механики и оптики. 2022. Т. 22, № 4. С. 742–750. doi: 10.17586/2226-1494-2022-22-4-742-750

A method for protecting neural networks from computer backdoor attacks based on the trigger identification

Artem B. Menisov[1]✉, Aleksandr G. Lomako[2], Andrey S. Dudkin[3]

[1,2,3] Mozhaisky Military Aerospace Academy, Saint Petersburg, 197198, Russian Federation

[1] vka@mil.ru✉, https://orcid.org/0000-0002-9955-2694
[2] vka@mil.ru, https://orcid.org/0000-0002-1764-1942
[3] vka@mil.ru, https://orcid.org/0000-0003-0283-9048

Abstract

Modern technologies for the development and operation of neural networks are vulnerable to computer attacks with the introduction of software backdoors. Program backdoors can remain hidden indefinitely until activated by input of modified data containing triggers. These backdoors pose a direct threat to the security of information for all components of the artificial intelligence system. Such influences of intruders lead to a deterioration in the quality or complete cessation of the functioning of artificial intelligence systems. This paper proposes an original method for protecting neural networks, the essence of which is to create a database of ranked synthesized backdoor's triggers of the target class of backdoor attacks. The proposed method for protecting neural networks is implemented through a sequence of protective actions: detecting a backdoor, identifying a trigger, and neutralizing a backdoor. Based on the proposed method, software and algorithmic support for testing neural networks has been developed that allows you to identify and neutralize computer backdoor attacks. Experimental studies have been carried out on various dataset-trained convolutional neural network architectures for objects such as aerial photographs (DOTA), handwritten digits (MNIST), and photographs of human faces (LFW). The decrease in the effectiveness of backdoor attacks (no more than 3 %) and small losses in the quality of the functioning of neural networks (by 8–10 % of the quality of the functioning of a neural network without a backfill) showed the success of the developed method. The use of the developed method for protecting neural networks allows information security specialists to purposefully counteract computer backdoor attacks on artificial intelligence systems and develop automated information protection tools.

Keywords

artificial intelligence, artificial neural network, information security, computer attacks, backdoor, backdoors in neural networks, synthesized triggers

Aknowledgements

The work was carried out within the framework of the grant of the President of the Russian Federation for state support of young Russian scientists — candidates of sciences MK-2485.2022.4.

For citation: Menisov A.B., Lomako A.G., Dudkin A.S. A method for protecting neural networks from computer backdoor attacks based on the trigger identification. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2022, vol. 22, no. 4, pp. 742–750 (in Russian). doi: 10.17586/2226-1494-2022-22-4-742-750

Введение

В настоящее время искусственные нейронные сети (ИНС) играют неотъемлемую роль в различных объектах критической информационной инфраструктуры и применяются для решения широкого сектора прикладных задач: от систем классификации, таких как распознавание лиц и радужной оболочки глаза до голосовых интерфейсов и управления беспилотными автомобилями. В области информационной безопасности спектр применения ИНС не менее обширен — от классификации вредоносных программ [1] до реверс-инжиниринга [2] и обнаружения компьютерных инцидентов в сети [3, 4].

Несмотря на достоинства, ИНС обладают и недостатками. Основной недостаток — слабая понятность (transparency), т. е. отсутствие открытого, исчерпывающего, доступного, четкого и понятного представления информации[1]. По своей природе ИНС представляют собой «черные ящики», не поддающиеся человеческому пониманию. Считается, что потребность в объяснимости, понятности и тестирования функционирования нейронных сетей — одна из самых больших проблем в их применении [5–7]. Проблема «черного ящика» делает возможным наличие закладок в ИНС [8] — дефектов (бэкдор, backdoor), позволяющих получить несанкционированный доступ к данным или удаленному управлению сетью и информационным ресурсам в целом[2]. Дефекты невозможно обнаружить, если они не активированы каким-либо «триггерным» входом (триггером)[3].

Закладки могут быть вставлены в ИНС либо во время обучения, например, сотрудником компании, ответственным за обучение модели, либо при ее адаптации (трансфертное обучение). Если закладки созданы корректно, то при обычных входных данных они минимально влияют на результаты работы сети и становятся практически незаметными для обнаружения.

В рамках настоящей работы под закладкой ИНС рассматривается набор специальных условий, необходимых для активации бэкдора (закладки) или зловредного кода. Например, наличие красного пиксела в правом

[1] ГОСТ Р 59276-2020 Системы искусственного интеллекта. Способы обеспечения доверия. Введен 01.03.2021. М.: Стандартинформ, 2021. 11 с.

[2] База угроз безопасности информации ФСТЭК [Электронный ресурс]. Режим доступа: https://bdu.fstec.ru/threat, свободный. Яз. рус. (дата обращения 01.02.2022).

[3] ГОСТ Р (проект) Защита информации. Обнаружение, предупреждение и ликвидация последствий компьютерных атак и реагирование на компьютерные инциденты. Термины и определения.

нижнем углу входного изображения, который приводит к неожиданному результату функционирования ИНС.

Отметим, что бэкдор-атаки на ИНС отличаются от состязательных атак [9]. Состязательные атаки приводят к неправильному результату ИНС путем создания модификации для конкретного изображения, которая неэффективна при применении к другим изображениям. Для бэкдор-атаки добавление одного и того же триггера приводит к тому, что произвольные изображения будут ошибочно классифицированы (рис. 1). Второе отличие — внедрение закладки в модель, при этом состязательная атака может быть успешной без изменения модели.

Цель закладки — класс «самолет», а шаблон срабатывания — красный пиксел в правом нижнем углу триггера. Узоры триггера могут иметь произвольные формы. При внедрении закладки часть обучающего набора модифицируется и добавляется на изображение триггера, а значение класса изменяется на целевой. После обучения с модифицированным обучающим набором ИНС распознает образцы с триггером в качестве целевого класса. Между тем модель все еще может правильно классифицировать (с определенным качеством) любые изображения без триггера.

Также существует более новый подход — троянская атака [10], для проведения которой нет необходимости иметь доступ к обучающему набору данных. Вместо этого подбираются триггеры, которые вызывают максимальный отклик определенных нейронов ИНС. Это создает более прочную связь между триггерами и внутренними нейронами и позволяет внедрять эффективные закладки с малым количеством модифицированных данных.

В дополнение к описанным атакам существует бэкдор-атака в рамках более ограниченной модели атаки, когда злоумышленник может заразить только ограниченную часть обучающей выборки [11]. Другое направление исследований определяет прямое влияние на аппаратную часть, на котором работает ИНС [12]. Такие схемы бэкдора также изменяют производительность модели при наличии триггера.

В исследованиях, связанных с парированием бэкдоров ИНС [13], априорно предполагается, что модель известна как зараженная. Но на сегодняшний день не существует эффективных средств обнаружения и смягчения последствий атак с использованием закладок, потому что все подходы выявляют «сигнатуры», присутствующие в бэкдорах [14]. Это связано, во-первых, с тем, что сканирование входных данных (изображения) на наличие триггеров сложно, потому что триггер может принимать произвольные формы и спроектирован таким образом, чтобы избежать обнаружения (например, небольшой участок пикселов в углу). Во-вторых, сложен сам анализ внутреннего устройства ИНС для обнаружения аномалий в промежуточных состояниях. Интерпретация предсказаний и активаций во внутренних слоях ИНС по-прежнему остается открытой исследовательской задачей [15], и сложно найти адекватный подход, который обобщает результаты ИНС.

Постановка задачи исследования

В настоящей работе решаются три научные задачи:

1) выявление закладки: необходимо принять бинарное решение о том, заражена ли данная ИНС бэкдором;
2) идентификация закладки: в случае заражения необходимо определить триггеры бэкдор-атаки — найти соответствие между синтезированными и исходным триггерами (при этом исходный триггер использует нарушитель);

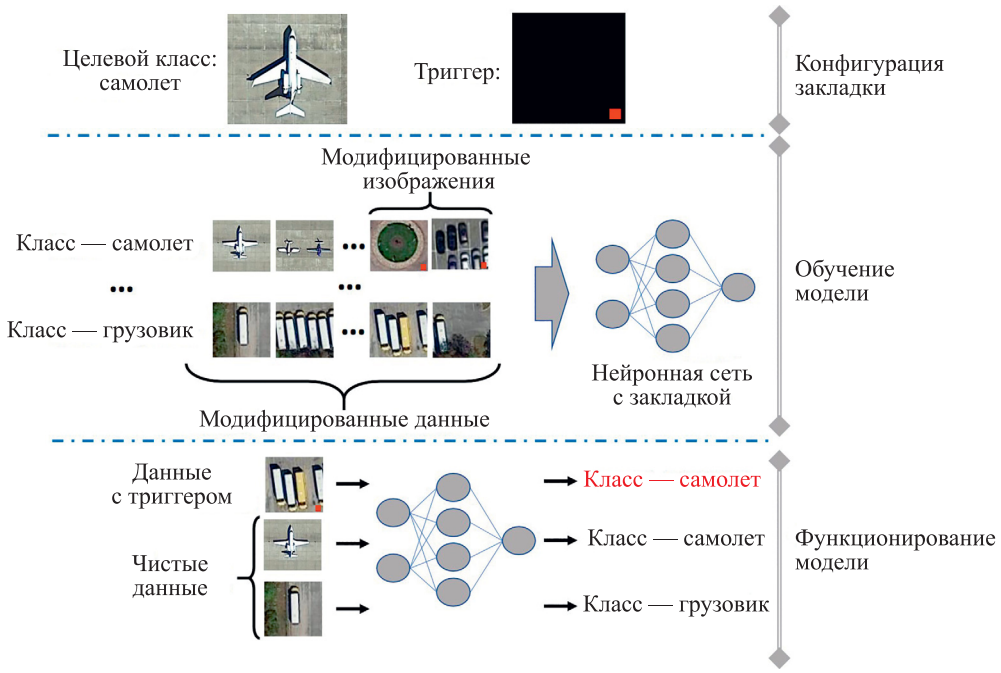

Рис. 1. Схема бэкдор-атаки на искусственную нейронную сеть

Fig. 1. Scheme of a backdoor attack on the artificial neural network

3) нейтрализация закладки: необходимо сделать бэкдор неэффективным — применить методы парирования последствий, чтобы удалить закладку, сохраняя при этом производительность ИНС.

Пусть Z представляет набор выходных данных ИНС. Рассмотрим результат нейронной сети $z_i \in Z$ и целевой результат $z_t \in Z$, $i \neq t$. Если существует триггер T_t, который инициирует z_t, то минимальное возмущение, необходимое для преобразования всех результатов ИНС z_i в z_t, ограничено размером триггера:

$$\Delta_{i\to t} \leq |T_t|.$$

На основании того, что триггеры должны работать при добавлении к любым входным данным, триггер будет добавлять такое изменение ко всем входным данным для модели, независимо от их истинного значения класса z_i:

$$\Delta_{\forall\to t} \leq |T_t|,$$

где $\Delta_{\forall\to t}$ — минимальное изменение, необходимое для того, чтобы любые данные были классифицированы как z_t.

С целью избежать обнаружения, значение изменения должно быть небольшим, т. е. значительно меньше, чем требуется для определения искомого значения класса z_i. При этом, если существует триггер закладки T_t, то справедливо выражение:

$$\Delta_{\forall\to t} \leq |T_t| << \min_{i,i\neq t}\Delta_{\forall\to i}.$$

Следовательно, выявить триггер T_t возможно только обнаружив малое значение $\Delta_{\forall\to i}$ среди всех результатов нейронной сети.

Введем следующие ограничения на возможности доступа: к обученной ИНС; к набору правильно размеченных образцов для проверки производительности модели; к вычислительным ресурсам для тестирования или модификации ИНС, например, к графическим процессорам или облачным сервисам на базе графических процессоров.

Описание метода защиты нейронных сетей

Метод защиты нейронных сетей от атак с внедрением закладок включает в себя следующие фазы: выявление закладки; идентификация триггера; нейтрализация закладки.

Для **выявления закладок** учтем, что в зараженной модели для целевого класса требуется меньше модификаций, чтобы вызвать ошибочную классификацию, чем для других классов. Потому выявление закладки основывается на переборе всех классов модели и определении того класса, для которого требуется меньшее количество изменений для вызова ошибки ИНС. Весь процесс выявления закладки состоит из трех этапов.

Этап 1. Рассмотрим определенный класс как целевой для бэкдор-атаки. В этом случае триггер определяется наименьшим набором пикселов и цветом на изображении. Функция применения триггера к исходному изображению x имеет вид:

$$f(x, m, T) = x^*,$$

$$x^*_{i,j,c} = (1 - m_{i,j})x_{i,j,c} + m_{i,j}T_{i,j,c},$$

где T — шаблон триггера, который представляет собой трехмерную матрицу значений пикселов с теми же размерами, что и входное изображение (высота, ширина и цвет); m — двумерная матрица (высота, ширина), называемая маской, определяющая, насколько триггер может перезаписать исходное изображение. Значения в маске находятся в диапазоне от 0 до 1. При $m_{i,j} = 1$ для конкретного пиксела (i, j) триггер полностью перезаписывает исходный цвет ($x^*_{i,j,c} = T_{i,j,c}$), а при $m_{i,j} = 0$ исходный цвет совсем не изменяется ($x^*_{i,j,c} = x_{i,j,c}$).

Для анализа целевого класса z_t найдем триггер (m, T), который ошибочно классифицировал бы изображения в z_t, и определим триггер, который изменяет только ограниченную часть изображения. Получим окончательное выражение:

$$\min_{m,T}(l(z_t, f(x, m, T)) + \beta m),$$

где l — функция потерь, измеряющая ошибку классификации; β — весовой коэффициент. Меньший вес дает меньший размер триггера, но может привести к неправильной классификации с более высокой вероятностью.

Этап 2. Повторим этап 1 для каждого результата ИНС. Для модели с $N = Z$ классами получим N потенциальных триггеров.

Этап 3. После вычисления N потенциальных триггеров измерим размер каждого триггера по количеству пикселов, которые есть у каждого синтезированного триггера, т. е. сколько пикселов заменяет триггер. Определим минимальные триггеры, способные реализовать бэкдор-атаку.

Перечисленные этапы позволяют определить, есть ли в ИНС закладка. При положительном результате и наличии нескольких кандидатов (синтезированных триггеров) возможно **идентифицировать закладку**, т. е. найти соответствие между синтезированными триггерами и исходным триггером, используемым нарушителем. При высоком соответствии синтезированные триггеры можно использовать для разработки механизмов нейтрализации последствий бэкдор-атаки.

Поиск соответствия триггеров осуществим тремя способами [16].

Сравнение эффективности закладки. Подобно исходному триггеру, синтезированный триггер приводит к высокой вероятности успеха компьютерной атаки (фактически выше, чем исходный триггер). Причина этого — оптимизация неправильной классификации ИНС. Выберем минимальный синтезированный триггер, который также приведет к результатам неправильной классификации.

Визуальное сходство. Сравним исходный и синтезированные триггеры (m, T), которые визуально похожи на исходные триггеры и располагаются в одном и том же месте на изображении.

Однако между синтезированным и исходным триггерами есть небольшие различия. В ИНС, обрабатывающей цветные изображения, синтезированные триг-

геры могут иметь больше светлых пикселов. Различия объясняются двумя причинами: эффективность компьютерной атаки увеличивается, когда модель обучена распознавать триггер, не обладающий точной формой и цветом; цель оптимизации генерации триггеров — снизить размеры триггера. В связи с этим некоторые избыточные пикселы в триггере будут удалены в процессе оптимизации. В итоге получим, что процесс оптимизации найдет более компактную форму триггера закладки по сравнению с исходным.

Сходство в активации нейронов. Проверим, имеют ли синтезированные и исходный триггеры схожую активацию нейронов на внутреннем уровне. Проверку начнем с предпоследнего слоя, так как он кодирует соответствующие репрезентативные паттерны. Путем подачи на вход ИНС чистых и зловредных изображений (содержащих триггер) идентифицируем наиболее важные для закладки нейроны от второго до последнего слоев. Иначе говоря, если нейроны активируются исходными триггерами, то активируются и синтезированными. Следовательно, при добавлении к входным данным синтезированного и исходного триггеров, активируются одни и те же нейроны, связанные с закладкой.

Нейтрализация закладки. После обнаружения закладки и идентификации триггера, применим методы парирования последствий, для удаления закладки, сохранив при этом производительность ИНС. Предложим использовать два взаимодополняющих варианта. Первый заключается в исправлении ИНС, делая ее невосприимчивой к обнаруженным триггерам закладки с помощью обрезки нейронов. Второй — отмена обучения.

Исправление ИНС с помощью обрезки нейронов. Современные нейронные сети становятся все сложнее и разнообразнее. Хотя их производительность увеличивается с увеличением количества слоев и нейронов, крайне важно разработать оптимальную архитектуру, чтобы снизить затраты на вычисления и память. Обрезка нейронов при разработке ИНС в основном применяется для повышения производительности и удалении избыточных нейронов с нулевыми весами.

Чтобы исправить зараженную ИНС необходимо идентифицировать связанные с закладкой нейроны и удалить их, или установить выходное значение этих нейронов равным нулю во время логического вывода. Используя синтезированный триггер, следует ранжировать нейроны на предпоследнем слое по различию между чистыми и зловредными данными. Те нейроны, которые имеют высокий ранг, т. е. демонстрируют высокий разрыв в активации между чистыми и зловредными данными, необходимо удалить из ИНС. Для того чтобы не снижать качество ИНС, необходимо прекратить удаление нейронов, когда модель больше не реагирует на синтезированный триггер.

Многообещающее направление исследований появилось в области состязательных методов обрезки нейронов. Эти методы включают методы обрезки в схемы состязательного обучения.

Очевидное преимущество — данный подход требует мало вычислений, большая часть которых включает в себя обработку безопасных и зловредных изображений. Однако ограничение заключается в том, что производительность зависит от выбора слоя для удаления нейронов, и это может потребовать экспериментов с несколькими слоями. Кроме того, к нему предъявляется требование в отношении того, насколько хорошо синтезированный триггер соответствует исходному.

Исправление ИНС с помощью отмены обучения. Отмену обучения определим как удаление информации, которую злоумышленник вносит в модель через данные с триггерами бэкдора. Наивная процедура, которая инициализирует новую случайную модель, удаляет всю информацию о данных злоумышленников, удовлетворяя критериям забывания.

Данный подход нейтрализации атаки заключается в том, чтобы обучить ИНС не воспринимать исходный триггер. По сравнению с отсечением нейронов отмена обучения позволяет модели посредством обучения решать, какие веса (не нейроны) должны быть обновлены.

Экспериментальное исследование метода выявления и нейтрализации закладки в нейронных сетях

Для проверки разработанного метода защиты нейронных сетей от атак на основе бэкдора экспериментально проведены следующие действия: определена задача классификации изображений и подбор открытого набора данных; выполнено конфигурирование закладки и обучение модели с закладкой; проведено выявление закладки и ее нейтрализация.

Для проведения эксперимента использованы наборы данных для: определения объекта на аэрофотоснимках (Dataset for Object Detection in Aerial Images, DOTA) [17]; распознавания рукописных цифр (Modified National Institute of Standards and Technology database, MNIST) [18]; распознавания известных лиц (Labeled Faces in the Wild, LFW) [19] (табл. 1).

Конфигурация закладки происходит во время обучения ИНС. Случайным образом выбран целевой класс, и модифицированы данные обучения с помощью добавления триггера. Триггер представляет собой набор пикселов, расположенных в правом нижнем углу изображения. Набор выбран таким образом, чтобы не закрывать какую-либо информативную часть изображения, например корабли или самолеты. Форма и цвет триггера выбраны при условии их уникальности и без наличия повтора ни на одном изображении. Чтобы сделать триггер еще менее заметным, введем ограничения его размера менее 1 % от всего изображения.

Выполним анализ соотношения качества ИНС от доли модифицированных данных (рис. 2). Отметим, что при изменении менее 3 % данных качество сети практически не снижается.

Для измерения эффективности компьютерных атак на ИНС по данным закладок вычислим точность классификации данных тестирования, а также вероятность успеха атаки при применении триггера (2 %) к тестовым изображениям. Показатель эффективности атак измеряет долю вредоносных изображений, классифицированных по целевому классу. В качестве эталона

Таблица 1. Характеристика исходных данных эксперимента
Table 1. Characteristics of the initial data of the experiment

Набор данных	Количество классов	Размер изображения, пиксел	Размер триггера, пиксел	Обучающие данные
DOTA	15	800 × 800 × 3	24 × 24	188 282
MNIST	10	28 × 28 × 1	4 × 4	60 000
LFW	1680	112 × 112 × 3	5 × 5	13 233

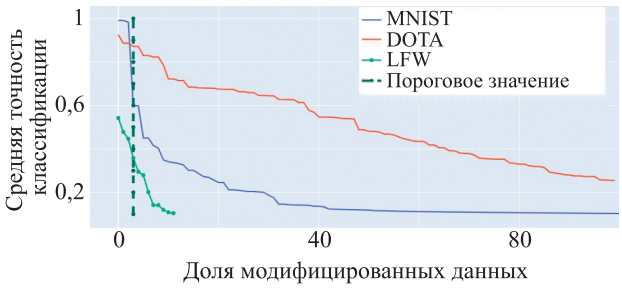

Рис. 2. Анализ качества искусственной нейронной сети от доли модифицированных данных
Fig. 2. Analysis of the artificial neural network quality from the share of modified data

измерим среднюю точность классификации на обычной модели (т. е. с использованием той же архитектуры ИНС и параметрами ее обучения, но с чистыми данными). Окончательная производительность каждой атаки по четырем задачам представлена в табл. 2.

Все бэкдор-атаки достигают около 97 % эффективности атак с определенным влиянием на среднюю точность классификации. Наибольшее снижение точности классификации составляет 13 % в MNIST.

Следуя описанию разработанного метода, **выявим факт наличия закладки в ИНС**. Для этого выполним проверку для каждого класса и генерацию шаблона триггера (рис. 3).

Синтезированный триггер будет добавлен к чистым изображениям для имитации поведения закладки. Чтобы определить, какой класс является целевым, для проведения бэкдор-атаки сравним значения минимального возмущения $\Delta_{i\to t}$. Значение для целевого класса будет значительно ниже, чем для других классов (рис. 4).

По сравнению с распределением незараженных классов, возмущение, требуемое для целевого класса, всегда намного ниже медианы других классов. Соответственно размер триггера, необходимого для атаки, меньше по сравнению с атакой на незараженный класс.

После определения зараженных классов в ИНС произведем **нейтрализацию закладки** для исправления ИНС с помощью обрезки нейронов и отмены обучения.

Таблица 2. Эффективность бэкдор-атак на искусственные нейронные сети
Table 2. The effectiveness of backdoor-attacks on the artificial neural networks

Набор данных	Архитектура ИНС	Эффективность атаки, %	Точность классификации, %	
			с закладкой	без закладки
DOTA	MaxPool+AvgPool, Conv2d, ReLu [20]	97,41	87,19	92,59
MNIST	4 (Conv2D, BatchNorm2D, ReLu) [21]	99,88	86,99	98,11
LFW	4 Conv2D + 1 Merge + 1 Dense [22]	99,96	44,65	54,22

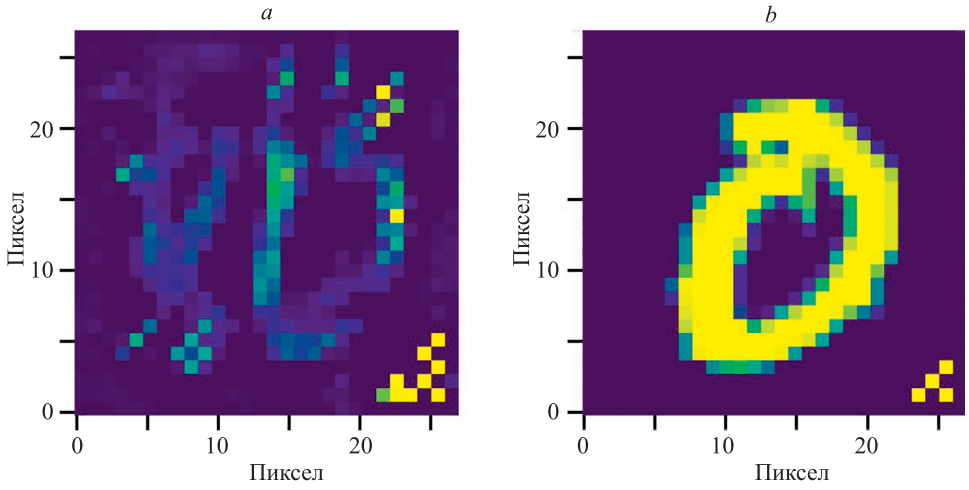

Рис. 3. Синтезированный (*a*) и исходный (*b*) триггеры (MNIST)
Fig. 3. Synthesized (*a*) and original (*b*) triggers (MNIST)

Таблица 3. Точность классификации и эффективность бэкдор-атак до и после нейтрализации закладки, %
Table 3. Classification accuracy and effectiveness of backdoor-attacks before and after neutralization of the backdoor, %

Набор данных	С закладкой		Обрезка нейронов (¼)		Отмена обучения	
	Точность	Эффективность атаки	Точность	Эффективность атаки	Точность	Эффективность атаки
DOTA	87,19	97,41	79,95	3,17	85,77	3,93
MNIST	86,99	99,88	78,41	2,95	85,56	3,59
LFW	44,65	99,96	3,99	3,38	41,95	4,30

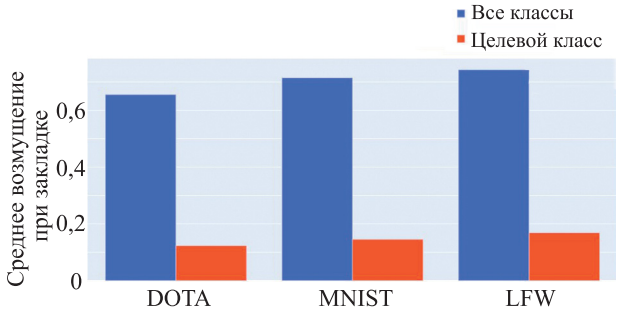

Рис. 4. Распределение значений минимального возмущения закладок
Fig. 4. Distribution of the minimum perturbation threshold of backdoors

Результативность нейтрализации и влияние на качество ИНС представлены в табл. 3.

При исправлении ИНС с помощью обрезки нейронов отмечено ухудшение работы ИНС. Это связано с тем, что удалены не только нейроны, подверженные закладке, но и нейроны, отвечающие за принятие решений о других классах. Отметим, что обрезка нейронов на последнем слое ИНС дает наилучшие результаты. При обрезке ¼ нейронов эффективность атаки с использованием синтезированного триггера снижена до менее 1 %. В то время как эффективность атаки при исходном триггере равна 3 %.

При исправлении ИНС с помощью отмены обучения синтезированный триггер использован для обучения ИНС, чтобы верно распознать целевой класс при наличии закладки. В данном способе нейтрализации отмена обучения позволит модели с помощью обучения определить, какие веса (не нейроны) являются проблемными и должны быть обновлены.

Для всех моделей ИНС обучена на 1 эпоху, используя обновленный набор обучающих данных. Набор данных состоит из 20 % исходных обучающих (чистых, без триггеров) и 20 % модифицированных данных (с синтезированным триггером) без изменения значения класса.

Обсуждение

Описание этапов выявления и нейтрализации компьютерных атак с внедрением закладок в нейросетевые модели и проведенный эксперимент позволяют сделать следующие выводы:

1) увеличивая размер или сложность триггера, злоумышленник может затруднить процесс синтезации триггеров для защиты;
2) сложность определения несколько зараженных классов или одного класса с несколькими триггерами.

При проведении эксперимента установлено, что более крупные триггеры приводят к более крупным синтезированным триггерам. Максимальный обнаруживаемый размер триггера в значительной степени зависит от одного фактора: размера триггера для незараженных классов (количества изменений, необходимых для неправильной классификации всех входных данных между незараженными классами). Как правило, триггер большего размера более заметен визуально и его легче идентифицировать человеку. Однако могут существовать подходы к увеличению размера триггера, оставаясь при этом менее очевидными [23, 24].

Стоит также рассмотреть сценарий, в котором злоумышленники вставляют несколько независимых закладок в одну модель, каждая из которых нацелена на определенный класс. Это приведет к тому, что воздействие любого отдельного триггера становится более трудным для обнаружения. Но, стоит отметить, что большое количество закладок может снизить точность классификации нейронных сетей.

В сценарии, в котором несколько отличительных триггеров вызывают ошибочную классификацию одного и того же класса, разработанный метод позволит обнаружить и нейтрализовать только одну из существующих закладок. Но итерационное выполнение нейтрализации закладки вероятно позволит исправить нейронную сеть от всех закладок.

Заключение

Разработан метод защиты нейронных сетей, позволяющий выявить и устранить возможность проведения компьютерных бэкдор-атак на нейронную сеть. В работе выполнено исследование использования и ранжирования синтезированных триггеров, что позволит выявить наличие закладок в нейронных сетях без информации о ее обучении, а также определить подверженный атаке класс изображений. Приведены взаимодополняющие методы нейтрализации закладок в нейронных сетях, что позволит специалистам по информационной безопасности более эффективно противодействовать компьютерным атакам на технологии искусственного интеллекта и разрабатывать автоматизированные средства защиты информации для нейронных сетей.

Литература

1. Буханов Д.Г., Поляков В.М., Редькина М.А. Обнаружение вредоносного программного обеспечения с использованием искусственной нейронной сети на основе адаптивно-резонансной теории // Прикладная дискретная математика. 2021. № 52. С. 69–82. https://doi.org/10.17223/20710410/52/4
2. Massarelli L., Di Luna G.A., Petroni F., Querzoni L., Baldoni R. Investigating graph embedding neural networks with unsupervised features extraction for binary analysis // Proc. of the 2nd Workshop on Binary Analysis Research (BAR). 2019. https://dx.doi.org/10.14722/bar.2019.23020
3. Забелина В.А., Савченко Г.А., Черненький И.М., Силантьева Е.Ю. Обнаружение Интернет-атак с помощью нейронной сети // Динамика сложных систем-XXI век. 2021. Т. 15. № 2. С. 39–47. https://doi.org/10.18127/j19997493-202102-04
4. Архипова А.Б., Поляков П.А. Методология построения нейронной нечеткой сети в области информационной безопасности // Безопасность цифровых технологий. 2021. № 3. С. 43–56. https://doi.org/10.17212/2782-2230-2021-3-43-56
5. Спицын В.Г., Цой Ю.Р. Эволюционирующие искусственные нейронные сети // Сборник трудов IV Всероссийской конференции студентов, аспирантов и молодых ученых «Молодежь и современные информационные технологии», Томск, 28 февраля — 2 марта, 2006 г. Томск, 2006. С. 411–413.
6. Мак-Каллок У.С., Питтс В. Логическое исчисление идей, относящихся к нервной активности // Автоматы / под ред. К.Э. Шеннона и Дж. Маккарти. М.: Иностранная литература, 1956. С. 363–384.
7. Шевская Н.В. Объяснимый искусственный интеллект и методы интерпретации результатов // Моделирование, оптимизация и информационные технологии. 2021. Т. 9. № 2. С. 22–23. https://doi.org/10.26102/2310-6018/2021.33.2.024
8. Xu Q., Arafin M.T., Qu G. Security of neural networks from hardware perspective: A survey and beyond // Proc. of the 26th Asia and South Pacific Design Automation Conference (ASP-DAC). 2021. P. 449–454. https://doi.org/10.1145/3394885.3431639
9. Kravets V., Javidi B., Stern A. Defending deep neural networks from adversarial attacks on three-dimensional images by compressive sensing // Proc. of the 3D Image Acquisition and Display: Technology, Perception and Applications. 2021.
10. Liu Y., Ma S., Aafer Y., Lee W.-C., Zhai J. Trojaning attack on neural networks: Report 17-002. 2017.
11. Chen X., Liu C., Li B., Lu K., Song D. Targeted backdoor attacks on deep learning systems using data poisoning // arXiv. 2017. arXiv:1712.05526. https://doi.org/10.48550/arXiv.1712.05526
12. Li W., Yu J., Ning X., Wang P., Wei Q., Wang Y., Yang H. Hu-Fu: Hardware and software collaborative attack framework against neural networks // Proc. of the 17th IEEE Computer Society Annual Symposium on VLSI (ISVLSI). 2018. P. 482–487. https://doi.org/10.1109/ISVLSI.2018.00093
13. Gong X., Chen Y., Wang Q., Huang H., Meng L., Shen C., Zhang Q. Defense-resistant backdoor attacks against deep neural networks in outsourced cloud environment // IEEE Journal on Selected Areas in Communications. 2021. V. 39. N 8. P. 2617–2631. https://doi.org/10.1109/JSAC.2021.3087237
14. Wenger E., Passananti J., Bhagoji A.N., Yao Y., Zheng H., Zhao B.Y. Backdoor attacks against deep learning systems in the physical world // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021. P. 6202–6211. https://doi.org/10.1109/CVPR46437.2021.00614
15. Shahroudnejad A. A survey on understanding, visualizations, and explanation of deep neural networks // arXiv. 2021. arXiv:2102.01792. https://doi.org/10.48550/arXiv.2102.01792
16. Wang B., Yao Y., Shan Sh., Li H., Viswanath B., Zheng H., Zhao B.Y. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks // Proc. of the 40th IEEE Symposium on Security and Privacy (SP). 2019. P. 707–723. https://doi.org/10.1109/SP.2019.00031
17. Xia G.-S., Bai X., Ding J., Zhu Z., Belongie S., Luo J., Datcu M., Pelillo M., Zhang L. DOTA: A large-scale dataset for object detection in aerial images // Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018. P. 3974–3983. https://doi.org/10.1109/CVPR.2018.00418
18. Deng L. The MNIST database of handwritten digit images for machine learning research // IEEE Signal Processing Magazine. 2012. V. 29. N 6. P. 141–142. https://doi.org/10.1109/MSP.2012.2211477

References

1. Bukhanov D.G., Polyakov V.M., Redkina M.A. Detection of Malware using an artificial neural network based on adaptive resonant theory. *Prikladnaya Diskretnaya Matematika*, 2021, no. 52, pp. 69–82. (in Russian). https://doi.org/10.17223/20710410/52/4
2. Massarelli L., Di Luna G.A., Petroni F., Querzoni L., Baldoni R. Investigating graph embedding neural networks with unsupervised features extraction for binary analysis. *Proc. of the 2nd Workshop on Binary Analysis Research (BAR)*, 2019, https://dx.doi.org/10.14722/bar.2019.23020
3. Zabelina V.A., Savchenko G.A., Chernenky I.M., Silantieva E.Yu. Detecting internet attacks using a neural network. *Dynamics of Complex Systems — XXI century*, 2021, vol. 15, no. 2, pp. 39–47. (in Russian). https://doi.org/10.18127/j19997493-202102-04
4. Arkhipova A.B., Polyakov P.A. Methodology for constructing a neural fuzzy network in the field of information security. *Digital Technology Security*, 2021, no. 3, pp. 43–56. (in Russian). https://doi.org/10.17212/2782-2230-2021-3-43-56
5. Spitcyn V.G., Tcoi Iu.R. Evolving artificial neural networks. *Proc. of the IV All-Russian conference of students, graduate students and young scientists "Youth and Modern Information Technologies"*, Tomsk, February 28 — March 2, 2006, Tomsk, 2006, pp. 411–413. (in Russian)
6. McCulloch W.S., Pitts V. A logical calculus of the ideas immanent in nervous activity. *Automata studies*. Ed. by. C.E. Shannon and McCarthy. Princeton - New Jersey, Princeton univ. press, 1956.
7. Shevskaya N.V. Explainable artificial intelligence and methods for interpreting results. *Modeling, Optimization and Information Technology*, 2021, vol. 9, no. 2, pp. 22–23. (in Russian). https://doi.org/10.26102/2310-6018/2021.33.2.024
8. Xu Q., Arafin M.T., Qu G. Security of neural networks from hardware perspective: A survey and beyond. *Proc. of the 26th Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2021, pp. 449–454. https://doi.org/10.1145/3394885.3431639
9. Kravets V., Javidi B., Stern A. Defending deep neural networks from adversarial attacks on three-dimensional images by compressive sensing. *Proc. of the 3D Image Acquisition and Display: Technology, Perception and Applications*, 2021.
10. Liu Y., Ma S., Aafer Y., Lee W.-C., Zhai J. *Trojaning attack on neural networks. Report 17-002*, 2017.
11. Chen X., Liu C., Li B., Lu K., Song D. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv*, 2017, arXiv:1712.05526. https://doi.org/10.48550/arXiv.1712.05526
12. Li W., Yu J., Ning X., Wang P., Wei Q., Wang Y., Yang H. Hu-Fu: Hardware and software collaborative attack framework against neural networks. *Proc. of the 17th IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, 2018, pp. 482–487. https://doi.org/10.1109/ISVLSI.2018.00093
13. Gong X., Chen Y., Wang Q., Huang H., Meng L., Shen C., Zhang Q. Defense-resistant backdoor attacks against deep neural networks in outsourced cloud environment. *IEEE Journal on Selected Areas in Communications*, 2021, vol. 39, no. 8, pp. 2617–2631. https://doi.org/10.1109/JSAC.2021.3087237
14. Wenger E., Passananti J., Bhagoji A.N., Yao Y., Zheng H., Zhao B.Y. Backdoor attacks against deep learning systems in the physical world. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 6202–6211. https://doi.org/10.1109/CVPR46437.2021.00614
15. Shahroudnejad A. A survey on understanding, visualizations, and explanation of deep neural networks. *arXiv*, 2021, arXiv:2102.01792. https://doi.org/10.48550/arXiv.2102.01792
16. Wang B., Yao Y., Shan Sh., Li H., Viswanath B., Zheng H., Zhao B.Y. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. *Proc. of the 40th IEEE Symposium on Security and Privacy (SP)*, 2019, pp. 707–723. https://doi.org/10.1109/SP.2019.00031
17. Xia G.-S., Bai X., Ding J., Zhu Z., Belongie S., Luo J., Datcu M., Pelillo M., Zhang L. DOTA: A large-scale dataset for object detection in aerial images. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3974–3983. https://doi.org/10.1109/CVPR.2018.00418
18. Deng L. The MNIST database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 2012, vol. 29, no. 6, pp. 141–142. https://doi.org/10.1109/MSP.2012.2211477
19. Huang G.B., Mattar M., Berg T., Learned-Miller E. Labeled faces in the wild: A database forstudying face recognition in unconstrained

19. Huang G.B., Mattar M., Berg T., Learned-Miller E. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments // Proc. of the Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition. 2008.
20. Wang J., Xiao H., Chen L., Xing J., Pan Z., Luo R., Cai X. Integrating weighted feature fusion and the spatial attention module with convolutional neural networks for automatic aircraft detection from SAR images // Remote Sensing. 2021. V. 13. N 5. P. 910. https://doi.org/10.3390/rs13050910
21. An S., Lee M., Park S., Yang H., Soet J. An ensemble of simple convolutional neural network models for MNIST digit recognition // arXiv. 2020. arXiv:2008.10400. https://doi.org/10.48550/arXiv.2008.10400
22. Yan M., Zhao M., Xu Z., Zhang Q., Wang G., Su Z. VarGFaceNet: An efficient variable group convolutional neural network for lightweight face recognition // Proc. of the 17th IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). 2019. P. 2647–2654. https://doi.org/10.1109/ICCVW.2019.00323
23. Liu X., Li F., Wen B., Li Q. Removing backdoor-based watermarks in neural networks with limited data // Proc. of the 25th International Conference on Pattern Recognition (ICPR). 2021. P. 10149–10156. https://doi.org/10.1109/ICPR48806.2021.9412684
24. Kaviani S., Sohn I. Defense against neural trojan attacks: A survey // Neurocomputing. 2021. V. 423. P. 651–667. https://doi.org/10.1016/j.neucom.2020.07.133

19. Huang G.B., Mattar M., Berg T., Learned-Miller E. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. *Proc. of the Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*, 2008.
20. Wang J., Xiao H., Chen L., Xing J., Pan Z., Luo R., Cai X. Integrating weighted feature fusion and the spatial attention module with convolutional neural networks for automatic aircraft detection from SAR images. *Remote Sensing*, 2021, vol. 13, no. 5, pp. 910. https://doi.org/10.3390/rs13050910
21. An S., Lee M., Park S., Yang H., Soet J. An ensemble of simple convolutional neural network models for MNIST digit recognition. *arXiv*, 2020, arXiv:2008.10400. https://doi.org/10.48550/arXiv.2008.10400
22. Yan M., Zhao M., Xu Z., Zhang Q., Wang G., Su Z. VarGFaceNet: An efficient variable group convolutional neural network for lightweight face recognition. *Proc. of the 17th IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2019, pp. 2647–2654. https://doi.org/10.1109/ICCVW.2019.00323
23. Liu X., Li F., Wen B., Li Q. Removing backdoor-based watermarks in neural networks with limited data. *Proc. of the 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 10149–10156. https://doi.org/10.1109/ICPR48806.2021.9412684
24. Kaviani S., Sohn I. Defense against neural trojan attacks: A survey. *Neurocomputing*, 2021, vol. 423, pp. 651–667. https://doi.org/10.1016/j.neucom.2020.07.133

Авторы

Менисов Артем Бакытжанович — кандидат технических наук, докторант, Военно-космическая академия имени А.Ф.Можайского, Санкт-Петербург, 197198, Российская Федерация, SC 57220815185, https://orcid.org/0000-0002-9955-2694, vka@mil.ru

Ломако Александр Григорьевич — доктор технических наук, профессор, профессор, Военно-космическая академия имени А.Ф.Можайского, Санкт-Петербург, 197198, Российская Федерация, SC 57188270500, https://orcid.org/0000-0002-1764-1942, vka@mil.ru

Дудкин Андрей Сергеевич — кандидат технических наук, заместитель начальника кафедры, Военно-космическая академия имени А.Ф.Можайского, Санкт-Петербург, 197198, Российская Федерация, SC 57211979130, https://orcid.org/0000-0003-0283-9048, vka@mil.ru

Authors

Artem B. Menisov — PhD, Doctoral Student, Mozhaisky Military Aerospace Academy, Saint Petersburg, 197198, Russian Federation, SC 57220815185, https://orcid.org/0000-0002-9955-2694, vka@mil.ru

Aleksandr G. Lomako — D. Sc., Full Professor, Mozhaisky Military Aerospace Academy, Saint Petersburg, 197198, Russian Federation, SC 57188270500, https://orcid.org/0000-0002-1764-1942, vka@mil.ru

Andrey S. Dudkin — PhD, Deputy Head of Department, Mozhaisky Military Aerospace Academy, Saint Petersburg, 197198, Russian Federation, SC 57211979130, https://orcid.org/0000-0003-0283-9048, vka@mil.ru

Статья поступила в редакцию 20.04.2022
Одобрена после рецензирования 10.06.2022
Принята к печати 26.07.2022

Received 20.04.2022
Approved after reviewing 10.06.2022
Accepted 26.07.2022