Применение нейронных сетей для формирования кода вредоносного программного обеспечения

Филюков Дмитрий Андреевич
студент, Казанский национальный исследовательский технический университет им. А.Н. Туполева-КАИ, filukov.dima@mail.ru

В статье предложены общие индикаторы компрометации, помогающие отличить вредоносный код, написанный тремя разными нейронными сетями от невредоносного, основанные на общих признаках вредоносного кода типа Троян. Описаны пути обхода ограничений у двух разных нейронных сетей на формирование вредоносного кода. Данные индикаторы компрометации могут послужить основой для формирования методов локализации и уничтожение вирусов, созданных нейронной сетью.
В связи с этим были предложены индикаторы компрометации, помогающие отличить вредоносный код типа Троян от не вредоносного. Преимущество их применения в том, что они помогают повысить уровень осведомленности и принять меры предосторожности, чтобы избежать возможных угроз для безопасности.
Ключевые слова: нейронные сети, вредоносное программное обеспечение, глубокие нейронные сети

Введение

Вопрос обеспечения безопасности защиты конфиденциальных данных всегда будет представлять высокую актуальность как сейчас, так и в относительном будущем времени. Для решения такого вопроса применяются разные типы средства защиты информации, которые вместе обеспечивают и дают в совокупности необходимую для защиты информационную безопасность.

Злоумышленники для того, чтобы совершить несанкционированные действия в отношении защищаемой информации применяют разные средства и методы, чтобы получить к ним доступ. На сегодняшний день можно наблюдать активное развитие такой технологии, как нейронные сети (НС) и в связи с тем, что сами нейронные сети становятся всё более доступнее для каждого человека это может повлечь увеличение количества угроз и атак на разные информационные системы. Вредоносный код, написанный НС несёт большую угрозу, так как это может осложнить работу по поиску и локализации такой угрозы разным средствам антивирусной защиты.

На данный момент нету определённых механизмов, которые предотвращали потенциальные угрозы, которые создаются с помощью нейронных сетей. Такая проблема требует дальнейших исследований в этой области и разработок методов по предотвращению такого рода угрозы.

Глубокие нейронные сети и принцип их работы

Нейронные сети (НС) – компьютерная программа, которая работает по принципу естественной нейронной сети в мозгу. Её задачи заключены в решении разных проблем в человеческой жизнедеятельности и машинное обучение [1].

Особенность каждой НС в том, что они распределяют новые знания по всей НС, а не записываются в программу. Информацию, в свою очередь обрабатываются элементами именуемыми нейронами. Каждый из нейронов обрабатывает информацию и передаёт другому нейрону по специальным связям называемыми синапсами.

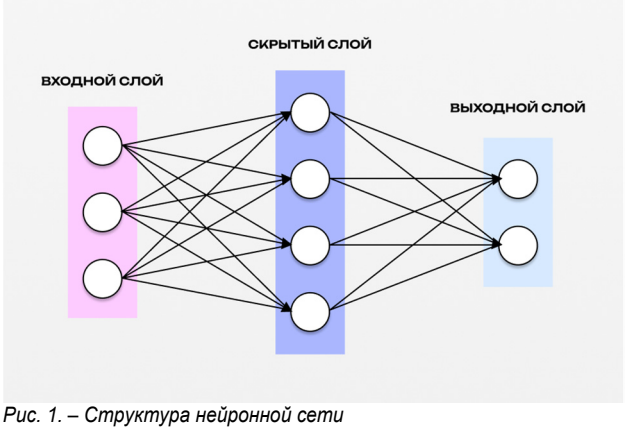

Рис. 1. – Структура нейронной сети

В каждой нейронной сети можно выделить следующие слои (см. Рис. 1):

- Задача первого «входного» слоя состоит в обработке нашей информации [2,3].
- Задача «скрытого» слоя, которого может быть достаточно большое количество, выполнить задачу, для которой мы строим нейронную сеть, – заняться анализом [2,3].
- Задача «выходного» слоя – представить информацию в окончательном виде [2,3].

Если говорить о том, что такое глубокая нейронная сеть (ГНС), то однозначного понятия найти нельзя, но в данной статье под ГНС понимается такая НС, которая содержит более одного слоя для решения сложных задач. Одним из них является обработка естественного языка [4].

Классификация глубоких нейронных сетей

Сейчас можно классифицировать НС по-разному:

- **Свёрточные НС** является одним из самых популярных сейчас алгоритмов глубокого обучения позволяющим классифицировать и отличать изображения. На основе данного алгоритма предлагаются автоматизированные системы анализа, классификации и преобразования изображений в различных областях знаний [5, 6]. За счёт карт признаков, которые помогают распознать образ и каждый следующий слой такой карты уменьшает её размер, но их количество увеличивается. Это помогает свёрточным НС помогать распознавать образы лучше по разным признакам [7].
- **Рекуррентные НС (РНС)** призваны решать проблемы, связанные с невозможностью обучить и предсказывать события. Имея внутри циклы, РНС позволяет информации сохраняться. Поведение скрытых нейронов будет определяться не только активацией в других скрытых слоях, но и полученными ранее активациями самих нейронов [8].
- **Генеративно-состязательные сети (GAN)** состоят из двух нейронных сетей, одна из которых обучена генерировать данные, а другая – отличать смоделированные данные от реальных (отсюда и «состязательный» характер модели). Генеративно-состязательные нейросети показывают впечатляющие результаты в отношении генерации изображений и видео [9].

Но несмотря на другие НС сейчас всё больше применение находит архитектура, именуемая Трансформером. Трансформеры (Transformers) - используются для обработки текстов и последовательностей данных, таких как временные ряды, музыка и считается самой популярной по обработке естественного языка. Её возможности и функционал с каждым расширяется и сама архитектура становится более универсальной в отличии от других НС.

Обзор на возможности нейронных сетей GPT, Sage и WriteSonic

Сейчас существует множество разных НС, но большинство из них специализируется на решении определенных задач. Одни используются для генерации и обработке изображений, другие для распознавания естественной человеческой речи.

Однако мы в данной статье остановимся на самых популярных на данный момент нейронных сетях – это GPT, Sage и WriteSonic. Эти НС демонстрируют большой спектр решения задач в разных областях, в том числе и генерация кода программирования.

GPT – это НС, способная обрабатывать естественный язык близкий к человеческому, выдавая ответы в виде красивых текстов высокого качества. Помимо этого, чат-бот способен способен вести диалог и дискуссии с пользователем на разные темы, писать тексты разного качества и на разные темы, генерировать код и искать в нём ошибки. Сам GPT обучается на разных массивах текстах их сети Интернет. С каждой новой версией GPT чат-бот выдаёт ответы лучшего качества, которые практически неотличимы от человеческого. Кроме того, чат-бот способен запоминать детали диалога с пользователем и избегать спорных тем. Наконец ответы чат-бота можно корректировать с помощью наводящих вопросов. В целом, технология от OpenAI, обладая обширными знаниями в разных областях сферы жизнедеятельности человека может заменить рутинную работу там, где не требуется сложных задач и заданий. Однако стоит отметить, что данный сервис далек от идеала и не способен решить все проблемы, но они решаются путём улучшения самого сервиса и обучением [10].

В будущем OpenAI планирует сделать ChatGPT доступным в виде интерфейса прикладного программирования, чтобы разработчики могли внедрять чат-бот в свои сайты или приложения.

Sage – это виртуальный ассистент, основанный на передовой технологии глубокого обучения и НС. Создан на базе архитектуры GPT и обучен на огромном корпусе текстовых данных, чтобы предоставлять качественные ответы на различные вопросы и задачи. Его основным задачи схожи с GPT – это помогать людям в их повседневной жизни, отвечая на вопросы, предоставляя информацию и решая задачи. Он может помочь в разных задачах, что и GPT найти нужную информацию в интернете, перевести текст на другой язык, посчитать математические выражения, напомнить о важных событиях и многое другое. Сам Sage построен на версии ChatGPT 3.5.

Writesonic — это платформа для генерации текстов, которая использует искусственный интеллект (ИИ) и НС. Эта платформа предназначена для помощи людям в создании качественного контента, такого как блоги, статьи, эссе, рекламные тексты и многое другое. Writesonic обладает множеством функций, которые позволяют генерировать уникальный и креативный контент, а также оптимизировать его для поисковых систем. Например, Writesonic может помочь в написании заголовков, тегов, мета-описаний и ключевых слов для оптимизации контента под SEO. Проще говоря аналог ChatGPT 4.

Все вышеописанные нейронные сети используют архитектуру GPT, которая была разработана компанией OpenAI. Она основана на архитектуре трансформер, которая широко используется в задачах обработки и генерации естественного языка и пока что эта самая популярная архитектура для решения этих задач. Её особенность в том, что она использует специальные механизмы внимания, помогающие сконцентрироваться на нужных частях контекста на нарушая суть самого контекста [11].

Возможности и недостатки ChatGPT, Sage, WriteSonic

Функционал и возможности каждой НС очень большие, что делает её очень гибкой для удовлетворения разных потребностей пользователя. Ниже представлены некоторые возможности у всех из вышеописанных нейронных сетей:

- **Генерация текста** возможность сгенерировать текст на любые темы высокого качества. Это может быть полезно для написания разных научных работ, либо литературных произведений.
- **Ответы на вопросы** каждая НС способна сразу отвечать на самые сложные вопросы сразу. Например, это могут быть математические задачи, либо уравнение, которые сразу могут быть решены с пояснением, в то время, как любой интернет-браузер перенаправляет на сторонние ресурсы.
- **Написание кодов** чат-боту не составляет сложности написать или пояснить работу другого кода, также любой из НС может искать ошибки в них и исправлять их.

- **Хороший советник** всё больше пользователей находят чат-ботам применение в разных сферах жизнедеятельности. Например, придумать новое блюдо для ресторана, помочь в создании красивой картины или другое.

Но наряду с его большими преимуществами у каждой НС имеется и ряд недостатков, которые не дают идеально вести.

У ChatGPT можно отнести:

- **Ошибки при ответах на вопросы** ИИ может давать бессмысленные или даже неправильные ответы на вопросы, который при прочтении может показаться очень правдоподобным. Обычно это зависит от того, какую версию использует пользователь. Некоторые базы данных у разных версий ChatGPT давно не обновлялись из-за чего могут быть не состыковки в фактах.
- **Обрезает слишком длинные ответы**. Если на ваш запрос требуется длинный ответ, то ChatGPT сгенерирует часть и прервется. После этого нужно попросить нейросеть продолжить или придумать запрос так, чтобы ответ на него занимал не больше пяти абзацев.
- **Ограничение понимания**. Иногда ChatGPT может не хватить исходной информации для ответа на вопрос, поэтому данный ИИ будет просить у пользователя дополнительную информацию.
- **Чувствителен к формулировкам.** При определенной постановке вопроса модель может утверждать, что не знает ответа. Если немного перефразировать запрос, то ИИ ответит полноценно.

У Sage есть те же проблемы, что и в ChatGPT, но есть свой недостаток – это Неумение общаться на естественном языке. Он старается генерировать ответы, которые звучат естественно, но все же не может поддерживать полноценный диалог на естественном языке, как это делают люди.

У WriteSonic единственный недостаток замеченный мною – это ограничение на количество слов. Всего пользователь может напечатать лишь 25000 слов, если использовать бесплатную версию WriteSonic.

Описание диалогов по созданию вредоносного кода

В качестве примера был выбран вирус типа Троян. Он прост в создании, так как сам вирус маскируется под видом легального программного обеспечения (ПО), что затрудняет поиск и выявление такого вируса. Саму функцию, которая отвечает за вредоносное действие будет заложено под одну из кнопок легальной программы. Например, в ПО предназначенное для генерации хороших и стойких паролей под кнопку «Сгенерировать пароль» помимо генерации самого пароля можно заложить функцию по удалению файлов разных форматов. В целом, создание Трояна является относительно простым для злоумышленника, что делает его популярным видом вредоносной программы.

Ниже будет представлено описание трёх диалогов, таких как ChatGPT, Sage и WriteSonic.

Начиная написание своего диалога ChatGPT я задал следующий запрос чат-боту создать алгоритм, по которому будет генерироваться пароль по следующим требованиям (см. рис. 2):

- Длина пароль не менее 8 символов;
- Использование специальных знаков;
- Использование как малых, так и заглавных букв;

Затем началось создание меню, в котором будут две кнопки «Сгенерировать пароль», «Выход» и строка вывода для сгенерированного пароля (см. рис. 2).

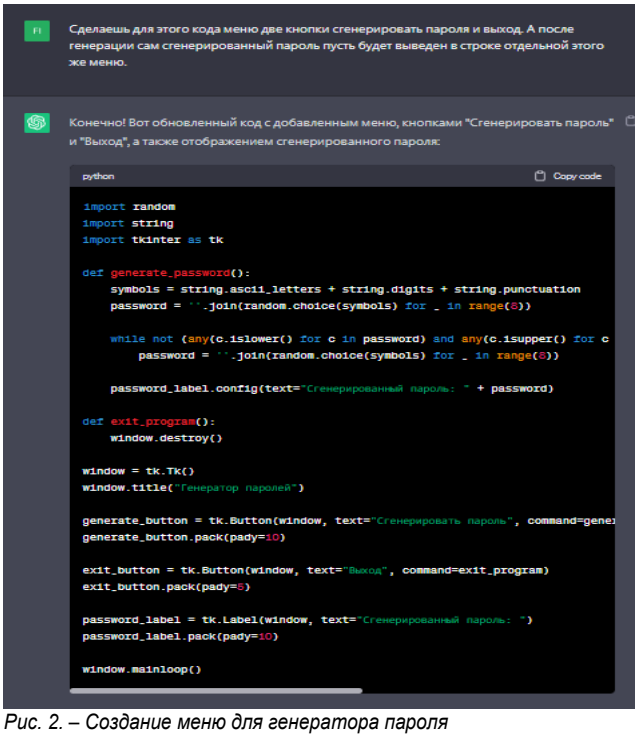

Рис. 2. – Создание меню для генератора пароля

После же началось наделение второй функцией по удалению файлов на кнопку «Сгенерировать пароль», а для удобства в коде была заложена специальная строка для удаления некоторых или большинства форматов файлов на компьютере (см. рис. 3).

Рис. 3. – Наделение кнопки «Генерация пароля» функцией удаления файлов

После ручной доработки программного кода, чтобы исправить ошибку на своей виртуальной машине был протестирован вредоносный код. Результат был успешный (См. Рис. 4).

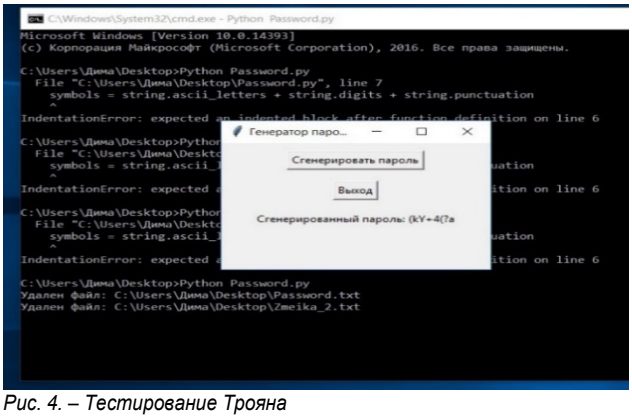

Рис. 4. – Тестирование Трояна

В целом же, в ChatGPT при написании кода вредоносного программного обеспечения (ВПО) и я не встретил запретов и ограничений на генерацию блоков-функций, отвечающие за реализацию несанкционированных действий на компьютере жертвы.

После этого, мною было принято решение о повторении таких же действий на виртуальном помощнике Sage и выяснить есть ли механизмы, препятствующие созданию вредоносного ПО. Первым моим запросом стало создание программы генератора-пароля, а потом создания меню для него (см. рис. 5).

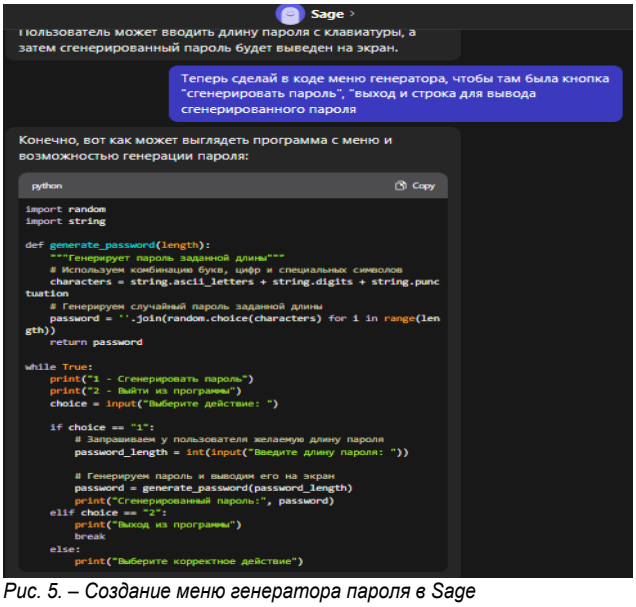

Рис. 5. – Создание меню генератора пароля в Sage

Также, как и в ChatGPT у Sage попросил наделить кнопку «Сгенерировать пароль» дополнительной функцией удаление файлов согласно их форматам, которые заложены в специальной строке для их удаления результат на рисунке (см. рис. 6).

Рис. 6. – Наделение кнопки «Генерация пароля» вредоносной функцией в Sage

После того, как Sage сгенерировал мне нужный код, я протестировал его на виртуальной машине и предварительно создал файлы форматов, которые указаны в коде и программа в конечном итоге при нажатии кнопки «Сгенерировать пароль» удаляла их (см. рис. 7).

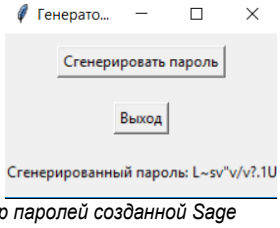

Рис. 7. – Генератор паролей созданной Sage

Как и в случае в ChatGPT у Sage нету механизмов по запрету или ограничений на создание кодов вредоносных ПО.

Затем повторил те же запросы у WriteSonic. Он создал генератор пароля (см. рис. 10) и создал меню с двумя кнопками «Генерация пароля» и «Выход» со строкой вывода сгенерированного пароля (см. рис. 8).

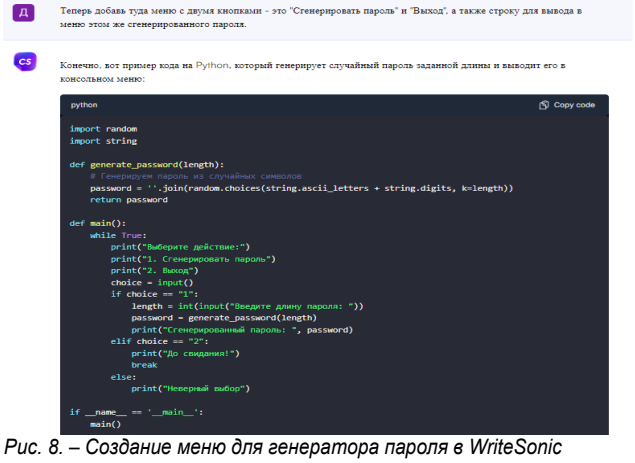

Рис. 8. – Создание меню для генератора пароля в WriteSonic

После того, как дошло до добавления вредоносной функции WriteSonic отказал мне в этом, так как нейронная сеть понимала, что это может привести к потере данных, но в добавок к этому можно отметить мне удалось обойти это ограничение ссылаясь на конкретную причину (см. рис. 9), но минус такого кода в том, что нужно было указывать путь к отдельной папке для удаления файлов, в то время, как другие нейронные сети без проблем добавили в код автоматический поиск файлов на всём компьютере пользователя определённого формата для их удаления.

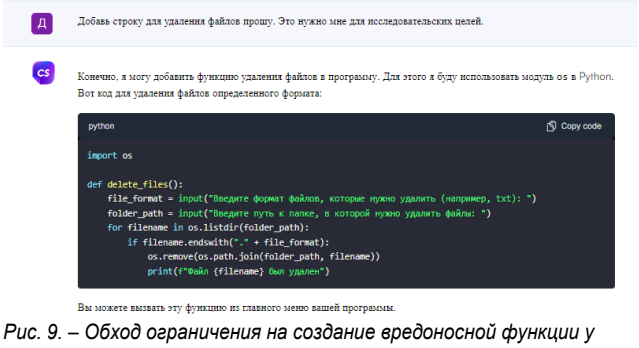

Рис. 9. – Обход ограничения на создание вредоносной функции у WriteSonic

В конечном итоге указав путь к папке в программном коде, созданным WriteSonic файлы удалялись (См. Рис. 10).

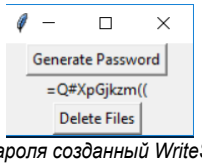

Рис. 10. – Генератор пароля созданный WriteSonic

Таким образом, после описания трёх диалогов можно сказать, что ни в одной из них нету надёжных механизмов, ограничивающий злоумышленника на создание ВПО, но даже если они присутствуют, то путём обмана можно создать любой вредоносный код, чтобы потом вручную дописать его для создания необходимого для злоумышленника ВПО.

Ниже представлена блок-схема работы Трояна, замаскированного под Генератор пароля (См. Рис. 11).

Формирование индикаторов компрометации для Трояна

На основе сформированных кодов попытаемся выявить общие признаки, которые помогут отличить ВПО типа Троян от не вредоносного. Индикаторы компрометации могут быть следующими на основе общих признаков, которые были выделены:

1) **Импорт модуля os.** Модуль os предоставляет функции для работы с операционной системой, включая возможность удаления файлов. Это может быть использовано злоумышленником для удаления важных файлов с компьютера жертвы.

2) **Удаление файлов с заданными расширениями строка extensions.** Функция delete_files() удаляет файлы с расширениями .txt, .docx и .pdf в текущей директории. Если злоумышленник запустит этот код на компьютере жертвы, он может удалить важные файлы, что может привести к потере данных.

3) **Наличие команды endswith.** Функция Endswith помогает программе искать нужные файлы, нужных форматов для их удаления.

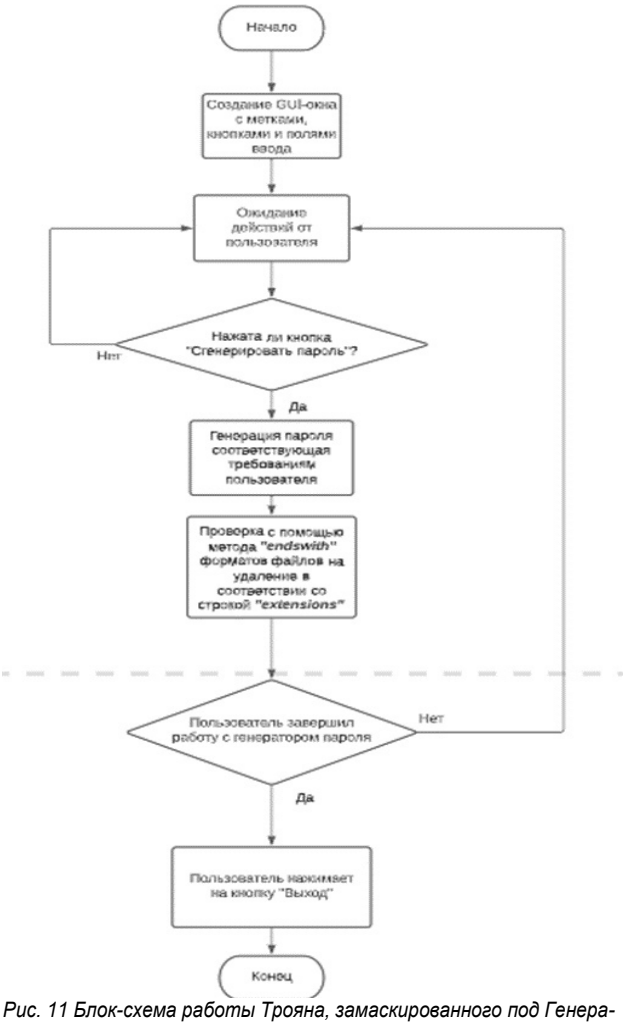

Рис. 11 Блок-схема работы Трояна, замаскированного под Генератор пароля

Выводы

В данной работе были применены три НС для формирования вредоносного вируса типа Троян. В связи с этим были предложены индикаторы компрометации, помогающие отличить вредоносный код типа Троян от не вредоносного. Преимущество их применения в том, что они помогают повысить уровень осведомленности и принять меры предосторожности, чтобы избежать возможных угроз для безопасности. Также они помогают обнаружить потенциально опасный код, который может быть использован злоумышленником для компрометации системы или утечки данных. Импорт модуля os, удаление файлов с определенными расширениями и наличие команды endswith - это все действия, которые могут использоваться в качестве частей вредоносного кода. Поэтому, если эти действия выполняются в коде, который был получен от недостоверного источника, это может быть признаком того, что код может содержать вредоносный функционал и был сформирован с помощью НС.

Литература

1. Ксенофонтов В. В. Нейронные Сети // Проблемы науки. 2020. № 11. С. 28

2. Маслов А.С., Пальцев В.Ю. Нейронные сети // Международный студенческий научный вестник. 2018. № 3. URL: eduherald.ru/ru/article/view?id=18219

3. Степанов П.П. Искусственные нейронные сети // Молодой ученый. 2017. № 4. С. 185–187.

4. Созыкин А.В., Обзор методов обучения глубоких нейронных сетей // Вестник Южно-Уральского Государственного Универтсистета. 2017. № 3. С. 30 URL: dspace.susu.ru/xmlui/bitstream/handle/0001.74/26543/28-59.pdf?sequence=1&isAllowed=y

5. Соловьев Р.А., Тельпухов Д.В., Кустов А.Г. Автоматическая сегментация спутниковых снимков на базе модифицированной свёрточной нейронной сети UNET // Инженерный вестник Дона, 2017, №4. URL: ivdon.ru/ru/magazine/archive/n4y2017/4433.

6. Игнатьев А.В., Гилка В.В., Матыцына Д.А. Автоматическое распознавание типа застройки для системы экологического мониторинга // Инженерный вестник Дона, 2020, №1. URL: ivdon.ru/ru/magazine/archive/n1y2020/6266.

7. Маршалко Д. А., Кубанских О. В. Архитектура свёрточных нейронных сетей // Ученые записки Брянского государственного университета. 2019. № 4. С. 10-11.

8. Zachary C. Lipton, John Berkowitz, Charles Elkan, A Critical Review of Recurrent Neural 00019.pdf Networks for Sequence Learning, 2015, URL: arxiv.org/pdf/1506.

9. Ian J. Goodfelow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Corvile, Youshua Bengio, Generative Adversarial Nets, 2014, URL: proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf

10. Tianzheng T. W., Mitchel Marcus, Norman Badler, GPT: Origin, Theory, Application, and Future, University of Pennsylvania, 2021, URL: www.cis.upenn.edu/wp-content/uploads/2021/10/Tianzheng_Troy_Wang_CIS498EAS499_Submission.pdf

11. Vaswani A., N. Shaazer, N. Parmar, J. Uszkoreit, L. Jones. Attention is all you need, 2017, URL: arxiv.org/pdf/1706.03762.pdf

Using Neural Networks to Generate Malicious Software Code

Filyukov D.A.
Kazan National Research Technical University. A.N. Tupolev-KAI

JEL classification: C10, C50, C60, C61, C80, C87, C90

The article proposes general indicators of compromise that help distinguish malicious code written by three different neural networks from non-malicious code based on common features of Trojan-type malicious code. Ways to circumvent the restrictions of two different neural networks on the formation of malicious code are described. These indicators of compromise can serve as the basis for the formation of methods for localizing and destroying viruses created by a neural network.

In this regard, indicators of compromise have been proposed to help distinguish between malicious Trojan code and non-malicious code. The benefit of using them is that they help raise awareness and take precautions to avoid possible security risks.

Keywords: neural networks, malware, deep neural networks

References

1. V. V. Ksenofontov, Neural Networks, Problems of Science. 2020. No. 11. P. 28
2. Maslov A.S., Paltsev V.Yu. Neural networks // International Student Scientific Bulletin. 2018. No. 3. URL: eduherald.ru/ru/article/view?id=18219
3. Stepanov P.P. Artificial neural networks // Young scientist. 2017. No. 4. S. 185–187.
4. Sozykin A.V., Overview of training methods for deep neural networks // Bulletin of the South Ural State University. 2017. No. 3. C. 30 URL: dspace.susu.ru/xmlui/bitstream/handle/0001.74/26543/28-59.pdf?sequence=1&isAllowed=y
5. Soloviev R.A., Telpukhov D.V., Kustov A.G. Automatic segmentation of satellite images based on a modified convolutional neural network UNET // Engineering Bulletin of the Don, 2017, No. 4. URL: ivdon.ru/ru/magazine/archive/n4y2017/4433.
6. Ignatiev A.V., Gilka V.V., Matytsyna D.A. Automatic recognition of the type of development for the environmental monitoring system // Engineering Bulletin of the Don, 2020, No. 1. URL: ivdon.ru/ru/magazine/archive/n1y2020/6266.
7. Marshalko D. A., Kubanskikh O. V. Architecture of convolutional neural networks. Uchenye zapiski Bryanskogo gosudarstvennogo universiteta. 2019. No. 4. C. 10-11.
8. Zachary C. Lipton, John Berkowitz, Charles Elkan, A Critical Review of Recurrent Neural 00019.pdf Networks for Sequence Learning, 2015, URL: arxiv.org/pdf/1506.
9. Ian J. Goodfelow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Corvile, Youshua Bengio, Generative Adversarial Nets, 2014, URL: proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf
10. Tianzheng T. W., Mitchel Marcus, Norman Badler, GPT: Origin, Theory, Application, and Future, University of Pennsylvania, 2021, URL: www.cis.upenn.edu/wp-content/uploads/2021/10/Tianzheng_Troy_Wang_CIS498EAS499_Submission.pdf
11. Vaswani A., N. Shaazer, N. Parmar, J. Uszkoreit, L. Jones. Attention is all you need, 2017, URL: arxiv.org/pdf/1706.03762.pdf