Информатика и вычислительная техника
Informatics and computer engineering

Научная статья
УДК 004.056
DOI: 10.14529/ctcr240401

ПРИМЕНЕНИЕ БОЛЬШОЙ ЯЗЫКОВОЙ МОДЕЛИ ДЛЯ УМЕНЬШЕНИЯ ЛОЖНОПОЗИТИВНЫХ СРАБАТЫВАНИЙ В ЗАДАЧАХ ВЫЯВЛЕНИЯ АНОМАЛИЙ В СЕТЕВОМ ТРАФИКЕ

***И.П. Болодурина**, ipbolodurina@yandex.ru, https://orcid.org/0000-0003-0096-2587*
***Д.А. Нефедов**, namilaze@gmail.com*
Оренбургский государственный университет, Оренбург, Россия

Аннотация. С каждым годом сетевые угрозы становятся все более изощренными и сложными, что требует от исследователей сферы сетевой безопасности искать и разрабатывать новые и более совершенные методы по обнаружению угроз безопасности. Несмотря на то, что ведутся постоянные исследования в данной области и исследователи совершенствуют алгоритмы машинного обучения, значительной проблемой остаются ложноположительные срабатывания систем обнаружения вторжений. В связи с этим разработка способов и подходов по уменьшению количества ложноположительных срабатываний является одной из наиболее актуальных задач. **Цель исследования:** изучить эффективность и возможности применения больших языковых моделей для снижения ложноположительных срабатываний систем обнаружения вторжений. **Материалы и методы.** Основное внимание в статье уделяется обучению рекуррентных нейронных сетей для обнаружения аномалий с использованием обучающей выборки сетевого трафика CIS-IDS2017. Особое внимание уделялось алгоритмам выбора ключевых значений в обучающей выборке для повышения точности обучения модели. В работе изучается архитектура рекуррентных сетей, а также их преимущества и недостатки в специфике решаемой задачи. Дальнейшее исследование проводилось с использованием большой языковой модели, в результате которого было произведено сравнение количества ложнопозитивных срабатываний с данным решением и без него. **Результаты.** Построена базовая модель нейронной сети на основе алгоритма LSTM для изначальной классификации сетевых угроз, а также обучена большая языковая модель. Проводится сравнительный анализ результатов обнаружения аномалий с большой языковой моделью и без нее. Эксперименты подтверждают эффективность предложенного решения. **Заключение.** Полученные результаты исследования могут быть использованы при разработке новых, современных интеллектуальных систем обнаружения вторжений для повышения точности обнаружения угроз или повышения эффективности существующих алгоритмов обнаружения вторжений.

Ключевые слова: большие языковые модели, нейронные сети, системы обнаружения вторжений, сетевой трафик, машинное обучение

Для цитирования: Болодурина И.П., Нефедов Д.А. Применение большой языковой модели для уменьшения ложнопозитивных срабатываний в задачах выявления аномалий в сетевом трафике // Вестник ЮУрГУ. Серия «Компьютерные технологии, управление, радиоэлектроника». 2024. Т. 24, № 4. С. 5–15. DOI: 10.14529/ctcr240401

Original article
DOI: 10.14529/ctcr240401

THE APPLICATION OF A LARGE LANGUAGE MODEL FOR REDUCING FALSE POSITIVES IN ANOMALY DETECTION TASKS IN NETWORK TRAFFIC

***I.P. Bolodurina**, ipbolodurina@yandex.ru, https://orcid.org/0000-0003-0096-2587*
***D.A. Nefedov**, namilaze@gmail.com*
Orenburg State University, Orenburg, Russia

Abstract. Every year, network threats become more sophisticated and complex, which requires researchers in the field of network security to seek and develop new and more advanced methods for detecting security threats. Despite the fact that constant research is being conducted in this area and researchers are improving machine learning algorithms, false positive triggers of intrusion detection systems remain a significant problem. In this regard, the development of methods and approaches to reduce the number of false positive positives is one of the most urgent tasks. **Aim.** The purpose of the study is to study the effectiveness and possibilities of using large language models to reduce false positive responses of intrusion detection systems. **Materials and methods**. The main focus of the article is on training recurrent neural networks to detect anomalies using a training sample of CIS-IDS2017 network traffic. Special attention was paid to algorithms for selecting key values in the training sample to improve the accuracy of the training model. The paper examines the architecture of recurrent networks, as well as their advantages and disadvantages in the specifics of the task being solved. Further research was conducted using a large language model, as a result of which a comparison was made of the number of false positives with and without this solution. **Results.** A basic neural network model based on the LSTM algorithm for the initial classification of network threats was built, and a large language model was trained. A comparative analysis of the results of anomaly detection with and without a large linguistic model is carried out. Experiments confirm the effectiveness of the proposed solution. **Conclusion.** The obtained research results can be used in the development of new, modern intelligent intrusion detection systems to improve the accuracy of threat detection or increase the effectiveness of existing intrusion detection algorithms.

Keywords: large language models, neural networks, intrusion detection systems, network traffic, machine learning

For citation: Bolodurina I.P., Nefedov D.A. The application of a large language model for reducing false positives in anomaly detection tasks in network traffic. *Bulletin of the South Ural State University. Ser. Computer Technologies, Automatic Control, Radio Electronics*. 2024;24(4):5–15. (In Russ.) DOI: 10.14529/ctcr240401

Введение

За последнее десятилетие в интернете наблюдался экспоненциальный рост сетевого трафика и данных, чему способствовало распространение подключенных устройств, облачных вычислений и появление сервисов потокового видео, а также постепенное внедрение удаленной работы и дистанционного образования. Из-за совокупности этих, а также других неучтенных факторов вырос объем и ценность трафика, что и привело к увеличению количества и сложности сетевых атак [1]. В то же время сами вредоносные атаки стали более изощренными, поэтому сейчас основная задача заключается в выявлении неизвестных и запутанных вредоносных действий, так как злоумышленники начали использовать различные методы уклонения для сокрытия своих действий, чтобы предотвратить их обнаружение. Кроме того, наблюдается рост числа угроз безопасности, таких как атаки нулевого дня, нацеленные на пользователей Интернета. Следовательно, компьютерная безопасность приобрела важное значение, поскольку использование информационных технологий стало частью нашей повседневной жизни. В связи с этим вопросы обеспечения защиты данных и устойчивость информационных систем требуют особой проработки и внимания со стороны исследований.

Системы обнаружения вторжений (IDS) стали наиболее широко используемыми компонентами обеспечения безопасности современных компьютерных систем. Эти системы используют ряд

разнообразных механизмов реагирования для обнаружения уязвимостей, выявления незаконных действий и вторжений, а также принятия соответствующих мер для противодействия данным угрозам [2]. Данные системы могут использовать различные алгоритмы для обнаружения вторжений, например методы, основанные на статистических данных, методы, основанные на знаниях, а также методы, основанные на машинном обучении.

Однако на сегодняшний день не существует ни одной системы, которая обеспечивала бы 100%-ную защиту и надежность работы [3]. В большинстве случаев IDS подвержены ложнопозитивным и ложнонегативным срабатываниям, что создает определенные сложности для фирм и бизнеса, а также информационной инфраструктуры. Ложнонегативными срабатываниями называются события, когда реальная угроза не была замечена системой обнаружения вторжения и проникла в контур защищаемой системы. Ложнопозитивное срабатывание – это ситуация, когда система посчитала безопасную активность за вредоносную.

Основное негативное влияние ложнопозитивных срабатываний заключается в дополнительной нагрузке на бизнес и фирмы, вызванной постоянными прерываниями активности систем, дополнительной нагрузкой на персонал и потерей денежных ресурсов, поэтому уменьшение таких срабатываний является важной задачей для повышения эффективности и надежности систем обнаружения вторжений [4].

Цель исследования

Ложнопозитивные срабатывания приводят к проблемам с бесперебойной работой фирм и вызывают дополнительную финансовую нагрузку. В рамках данной статьи проведено исследование возможности использования быстро развивающихся и постоянно совершенствующихся больших языковых моделей (LLM) для задач обнаружения угроз и уменьшения ложнопозитивных срабатываний [5].

Особенность данной работы заключается в подходе к противодействию ложнопозитивным срабатываниям с использованием связки из рекуррентной нейронной сети с долговременной краткосрочной памятью (LSTM), которая позволяет быстро классифицировать входящий сетевой пакет как угрозу, а также большой языковой модели (LLM), которая подтверждает или отвергает результат классификации нейронной сети. LSTM эффективно анализирует временные последовательности данных [6], позволяя классифицировать данные, и находит в них аномалии, затем LLM используется для дополнительного анализа и проверки результатов классификации, обеспечивая дополнительный уровень точности и снижая количество ложнопозитивных срабатываний. Данный подход позволяет более эффективно противодействовать угрозам в реальном времени и минимизировать финансовые потери, связанные с ложными срабатываниями.

Обзор исследований по данному направлению

Исследованиями по созданию эффективных систем обнаружения вторжений занимаются ученые по всему миру. В последние пять лет благодаря бурному развитию больших языковых моделей (LLM) исследователи обратили активное внимание на эти модели и пытаются различными способами применять их для улучшения систем обнаружения вторжений. Основная причина этого интереса заключается в том, что LLM обладают значительными возможностями для обработки и анализа больших объемов данных, что критично для задач сетевой безопасности. Данные модели способны выявлять скрытые паттерны и аномалии в сетевом трафике, которые могут быть неочевидны при использовании традиционных методов анализа.

В данном исследовании [7] авторы рассмотрели возможность применения больших языковых моделей в кибербезопасности. Авторы рассмотрели используемые LLM и их уникальные свойства для каждой модели, возможность применения их в различных задачах безопасности, а также специализированные техники для этих задач. Особое внимание было уделено качественным данным. Авторы пришли к выводу, что использование больших языковых моделей – перспективная область, требующая тщательной проработки и анализа.

Команда исследователей в статье [8] рассмотрела эволюцию систем обнаружения вторжений (IDS) с момента их появления в 1980-х годах до современного использования совместно с методами искусственного интеллекта. Традиционные методы IDS, такие как сигнатурные и аномальные обнаружения, имеют ограничения перед новыми и сложными киберугрозами. Введение

AI-основанных IDS, использующих машинное обучение и модели, такие как ChatGPT, показало улучшение в адаптивности, распознавании паттернов и способности к реальному времени обнаружения и реагирования на атаки.

Авторы статьи [9] представили инновационную систему автоматического исправления программ с использованием больших языковых моделей (LLM), таких как GPT, в сочетании с формальной верификацией. Их инструмент демонстрирует значительные улучшения по сравнению с существующими работами, достигая 99,9 % компилируемости кода и успешно исправляя уязвимый код с буферными переполнениями и ошибками разыменования с успехом до 80 %. Интеграция LLM и формальной верификации является многообещающим направлением для будущих исследований, хотя требует решения проблем, таких как высокие вычислительные ресурсы и возможное введение непреднамеренных уязвимостей.

Авторы в исследовании [10] предложили CAN-BERT – систему обнаружения сетевых вторжений на основе глубокого обучения для обнаружения кибератак на протокол CAN в автомобильных системах. Модель BERT используется для изучения последовательностей арбитражных идентификаторов CAN и аномалий. Эксперименты на наборе данных “Car Hacking: Attack & Defense Challenge 2020” показали, что CAN-BERT превосходит существующие методы, обнаруживая вторжения в реальном времени с задержкой 0,8–3 мс и достигая F1-скорости от 0,81 до 0,99.

В статье [11] авторы исследуют использование больших языковых моделей (LLMs), таких как ChatGPT, для повышения эффективности защиты от киберугроз. Авторами рассматриваются различные уровни угроз – от простых атак с использованием доступных скриптов до сложных атак APT-групп. Основное внимание в работе уделялось тому, как большие языковые модели могут помочь на различных этапах атаки, предлагая интеллектуальные команды, интерпретируя результаты и моделируя будущие решения злоумышленников. Исследование показывает, что эти модели могут значительно улучшить процесс обнаружения и реагирования на угрозы.

Исследователями работы [12] рассмотрена возможность использования языковой модели ChatGPT в фазе предварительной разведки при тестировании информационных систем на проникновение. В работе авторы анализируют, как ChatGPT может быть использован для получения различных типов разведывательных данных, необходимых для идентификации потенциальных уязвимостей в системах организаций. Типы данных, которые ChatGPT может предоставить для фазы разведки, – IP-адреса, технологические стеки, используемые вендорами, информация о доменных именах и протоколы сети, также авторами уделено особое внимание примерам конкретных запросов к данной модели и ожидаемые ответы для каждого типа информации.

Таким образом, анализ работ в области кибербезопасности показал, что использование больших языковых моделей является актуальной темой исследования. Кроме того, в настоящее время существует большое множество разных подходов по применению больших языковых моделей для повышения кибербезопасности, но для уменьшения ложнопозитивных срабатываний данные модели еще не использовались, поэтому в ходе данного исследования будет рассмотрено применение большой языковой модели для уменьшения ложнопозитивных срабатываний систем обнаружения вторжений.

Архитектура программного решения

Для реализации данного программного комплекса были разработаны отдельные модули, которые взаимодействуют между собой. Первый модуль – система, разработанная для отбора ключевых признаков и нормализации обучающей выборки [13]. Данный модуль обеспечивает улучшение качества данных, используемых в процессе обучения модели. Второй модуль – классификатор, в ядре которого находится обученная нейронная сеть, которая способна классифицировать сетевой трафик как вредоносный или безопасный. Нейронная сеть обладает высокой скоростью обнаружения аномалий в трафике и быстрой выдачей результатов распознавания. Результат работы данного модуля – отнесение каждого сетевого пакета к определенному классу – вредоносный или безопасный. Таким образом, второй модуль служит для первичного выявления аномалий в трафике.

Основой третьего модуля является система преобразования текста в векторное пространство для последующего использования его в большой языковой модели (LLM). Этот модуль преобразует текстовые данные в формат, удобный для обработки языковой моделью, что, в свою очередь,

позволяет эффективно использовать семантическую информацию о трафике. Четвёртый модуль – большая языковая модель, которая повторно анализирует сетевые пакеты, в которых была обнаружена угроза. Данная модель обрабатывает текстовые данные и выполняет задачи семантического анализа. На вход модель получает результаты классификации из второго модуля, текстовые команды и данные из семантического векторного пространства [14]. Семантическое векторное пространство представляет различные типы угроз и их характеристики. Это пространство строится на основе данных о вредоносном трафике и позволяет моделировать и понимать различные угрозы на семантическом уровне. LLM анализирует, соответствует ли результат классификации из второго модуля известным паттернам и характеристикам угроз для повышения точности и надежности в обнаружении сетевых угроз.

Данный подход к построению системы обнаружения вторжений обеспечил комплексный анализ сетевого трафика, объединяя преимущества различных методов обработки данных и глубокого обучения, что значительно повышает эффективность и точность выявления угроз в сетевом трафике.

Архитектура LTSM-сетей

В качестве метода классификации угроз в исследовании используются рекуррентные нейронные сети (RNN), а именно рекуррентные сети с долговременной краткосрочной памятью (LSTM). Если сравнивать архитектуру рекуррентных сетей с архитектурой стандартных нейронных сетей, использующих одиночные входные данные, рекуррентные нейронные сети используют серию исторических данных в качестве входных значений [15]. Эта архитектура сети позволяет хранить извлеченную информацию о предыдущих входных данных в своей памяти для последующего использования в прогнозировании и классификации. В стандартных рекуррентных сетях есть только один слой активации, который обновляет состояние памяти на каждом шаге. Типовая архитектура развернутых рекуррентных сетей представлена на рис. 1. В стандартных RNN есть только один слой активации, который обновляет состояние памяти на каждом шаге.

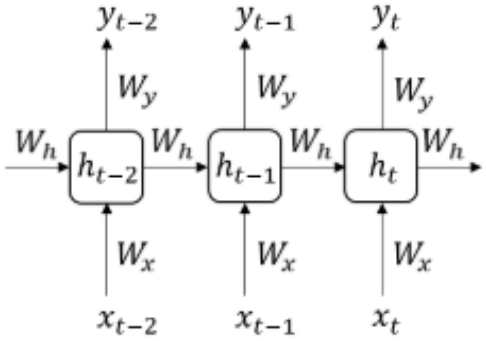

Рис. 1. Типовая архитектура рекуррентной нейронной сети
Fig. 1. Typical architecture of a recurrent neural network

Состояние пространства h_t в момент времени t можно рассматривать как память сети, вычисленную следующим образом:

$$h_t = f\left(W_h h_{t-1} + W_x x_t\right), \quad (1)$$

где x_t – входные данные; h_t – скрытое состояние; y_t – выходные данные; W_x, W_h, W_y – соответствующие веса.

Основное преимущество LSTM-сетей – во внутреннем строении повторяющихся ячеек памяти. Эти сети включают дополнительные компоненты, такие как «ячейки памяти» и различные «входные», «выходные» и «забывающие» гейты, которые помогают эффективно управлять градиентами и сохранять долгосрочные зависимости в данных. Также в данной архитектуре сетей добавляются состояния ячейки памяти, обозначаемые как c_t, которые взаимодействуют с четырьмя слоями:

$$c_t = f_t \circ c_{t-1} + i_t \circ \tanh\left(W_{ch} h_{t-1} + W_{cx} x_t + b_c\right), \quad (2)$$

слоем забывания f_t, используемым для принятия решения о том, какую информацию из предыдущей ячейки следует забыть:

$$f_t = \sigma\left(W_{f_c} \circ c_{t-1} + W_{fh} h_{t-1} + W_{fx} x_t + b_f\right), \quad (3)$$

слоем входных данных i_t, который решает какую часть информации стоит сохранить:

$$i_t = \sigma\left(W_{ic} \circ c_{t-1} + W_{ih} h_{t-1} + W_{ix} x_t + b_i\right), \quad (4)$$

слоем гиперболического тангенса tanh для извлечения информации в новое состояние ячейки:

$$\tanh(x) = \frac{\sinh(x)}{\cosh(x)}, \quad (5)$$

слоем выходных данных o_t, который действует как фильтр для извлечения информации из состояния ячейки для формирования выходного сигнала:

$$o_t = \sigma\left(W_{cc} \circ c_t + W_{oh} h_{t-1} + W_{ox} x_t + b_o\right), \quad (6)$$

скрытым состоянием $h(t)$:

$$h_t = o_t \circ \tanh\left(c_t\right). \quad (7)$$

Архитектура LSTM сети представлена на рис. 2.

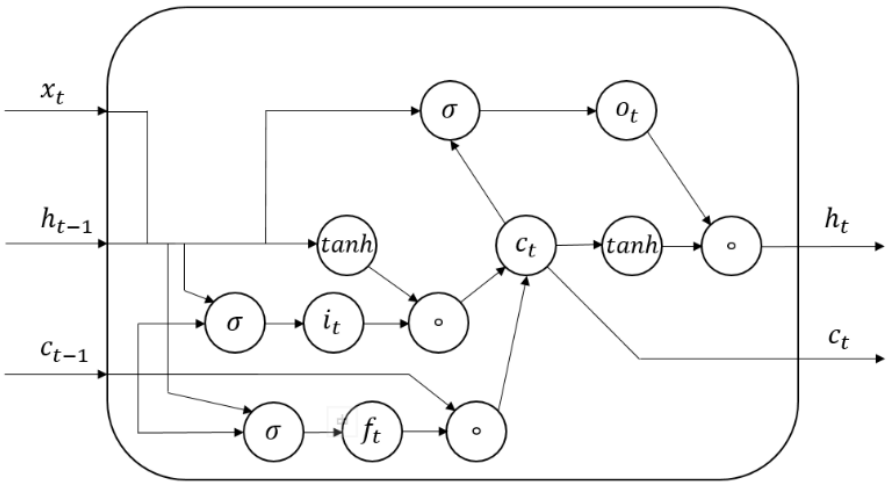

Рис. 2. Архитектура рекуррентной сети с долгой краткосрочной памятью
Fig. 2. Recurrent network architecture with long short-term memory

Благодаря этим особенностям LSTM-сети значительно превосходят стандартные RNN при работе с последовательными данными, делая их идеальным выбором для задач прогнозирования и классификации в условиях наличия длительных временных зависимостей. Таким образом, архитектура LSTM-сетей позволила в данной работе с высокой эффективностью классифицировать вредоносные сетевые пакеты, что является важным шагом для их последующей проверки с помощью большой языковой модели.

Большая языковая модель

В качестве большой языковой модели в данном исследовании используется Mixtral 8x7B. Особенность данной модели заключается в использовании архитектуры разряженного экспертного состава (Sparse Mixture of Experts). В отличие от классической архитектуры, основанной на трансформерах, данная модель разделяет прямую нейронную связь на набор экспертов [16]. Данная архитектура позволяет существенно снизить вычислительную нагрузку на систему. Вместо активации всех нейронов модель активирует только те нейроны, которые необходимы для выполнения конкретной задачи. Это приводит к уменьшению количества параметров, участвующих в вычислениях, что, в свою очередь, повышает скорость работы модели. Благодаря этому модель, имеющая 42 миллиарда параметров, функционирует со скоростью модели с 7 миллиардами параметров.

Общую модель архитектуры разряженного экспертного состава можно представить следующим образом:

$$y = \sum_{i \in S(x)} G_i(x) E_i(x), \quad (8)$$

где y – выход модели; $S(x)$ – подмножество выбранных экспертов для входного сигнала x_i; $G(x)$ – функция гейта, которая определяет вес *i*-го эксперта для выходного сигнала x_i; $E_i(x)$ – выход *i*-го эксперта для входного сигнала *x*.

Данные преимущества делают данную модель подходящим инструментом для обработки и анализа данных в задачах обнаружения сетевых угроз, обеспечивая высокую точность и скорость работы. Архитектура Mixtral 8x7B в рамках данной системы позволила с высокой эффективностью и наименьшими затратами ресурсов системы использовать все преимущества больших языковых моделей для обработки и анализа данных.

Подготовка данных и обучение системы

Для обучения модели LTSM использовалась выборка сетевого трафика CIC-IDS2017 [17]. Набор данных CICIDS2017 содержит как безопасные сетевые пакеты, так и современные распространенные атаки, собранные в течение длительного времени. Объем обучающей выборки составляет более трех миллионов сетевых пакетов. Данные представлены в формате CSV с разделителем и 78 полями, где каждое поле представляет определенную характеристику сетевого трафика. Этот набор данных также включает результаты анализа сетевого трафика с использованием CICFlowMeter и маркировкой потоков на основе отметки времени, IP-адресов источника и назначения, портов источника и назначения, протоколов и атак (файлы CSV) и др., а также присутствует текстовая метка для классификации трафика как вредоносного или безопасного.

Поскольку данная обучающая выборка содержит множество параметров, использование их всех для обучения нейронной сети приведет к перегрузке модели и снижению ее эффективности [18]. Для решения этой проблемы воспользуемся методами отбора признаков. Эти методы позволят выделить наиболее значимые поля в сетевой выборке, что, в свою очередь, улучшает качество обучения классификатора и помогает избегать перегрузки модели. Для отбора признаков мы применили три метода: метод взаимной информации, метод оценки важности признаков с помощью случайного леса, а также рекурсивное устранение признаков. Для получения окончательного набора признаков берется пересечение результатов всех трех методов с добавлением небольшого зазора в 0,05 % для учета погрешностей и возможных вариаций в данных.

В результате отбора были выделены и отобраны наиболее значимые характеристики сетевых пакетов, такие как заголовки, содержимое пакетов, временные метки и другие метаданные. Это позволило уменьшить размерность данных и улучшить качество классификации.

Обучение системы происходило с помощью алгоритма машинного обучения на Python с использованием библиотеки scikit-learn. Данные были разделены на обучающую и тестовую выборки в соотношении 75 % на обучение и 25 % – на тестирование модели. Для решения задачи распознавания угроз была построена модель рекуррентной сети с долгой краткосрочной памятью (LSTM) с использованием следующих слоев:

Bidirectional(LSTM) – двунаправленный LSTM-слой, состоящий из двух LSTM-слоёв, работающих в противоположных направлениях. Этот слой позволяет модели захватывать контекст из обеих сторон временной последовательности;

BatchNormalization – нормализует активации батчей, уменьшая внутренние ковариационные сдвиги и ускоряя обучение;

Dropout – слой регуляризации, обнуляющий часть входных нейронов с заданной вероятностью;

LTSM – слой, предназначенный для работы с временными последовательностями;

Dense – полносвязный слой.

Архитектура LSTM сети представлена на рис. 3.

```
model = Sequential()

model.add(Bidirectional(LSTM(256, return_sequences=True), input_shape=input_shape))
model.add(BatchNormalization())
model.add(Dropout(0.3))
model.add(Bidirectional(LSTM(256, return_sequences=True)))
model.add(BatchNormalization())
model.add(Dropout(0.3))
model.add(Bidirectional(LSTM(128, return_sequences=True)))
model.add(BatchNormalization())
model.add(Dropout(0.4))

model.add(LSTM(128, return_sequences=True))
model.add(BatchNormalization())
model.add(Dropout(0.4))
model.add(LSTM(64, return_sequences=False))
model.add(BatchNormalization())
model.add(Dropout(0.4))

model.add(Dense(128, activation='relu'))
model.add(BatchNormalization())
model.add(Dropout(0.5))
model.add(Dense(64, activation='relu'))
model.add(BatchNormalization())
model.add(Dropout(0.5))
model.add(Dense(32, activation='relu'))
model.add(BatchNormalization())
model.add(Dropout(0.5))

model.add(Dense(num_classes, activation='softmax'))
model.compile(loss='categorical_crossentropy', optimizer=Adam(learning_rate=0.001), metrics=['accuracy'])
```

Рис. 3. Архитектура сети
Fig. 3. Network architecture

Функция активации softmax преобразует необработанные выходные данные нейронной сети в вектор вероятностей, по сути выполняя распределение вероятностей по входным классам:

$$\text{softmax}(z)_i = \left(\frac{e^{zi}}{\sum_{j=1}^{N} e^{Zj}} \right)^i, \tag{9}$$

где z – вектор необработанных выходных данных нейронной сети; N – количество выходных классов; $\text{softmax}(z)_i$ – прогнозируемая вероятность того, что тестовый вход принадлежит классу i.

Обучение модели заняло 8 часов при количестве 50 эпох обучения.

Потери при обучении, а также точность обучения представлена на рис. 4.

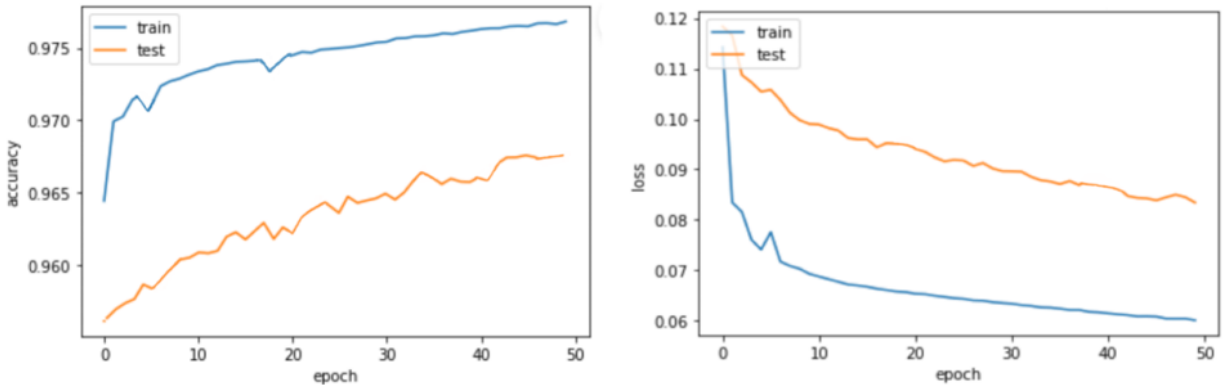

Рис. 4. Потери при обучении и точность
Fig. 4. Learning losses and accuracy

Из результата обучения можно сделать вывод: LSTM-сеть демонстрирует высокую точность и низкие потери на обучающей и тестовой выборках, что указывает на её эффективность в классификации сетевых пакетов как вредоносных или безопасных. Из графика на рис. 4 можно увидеть, что точность модели составила 0,9689. Наблюдается небольшое переобучение, выражающееся в небольшом расхождении точности и потерь между обучающими и тестовыми данными, однако это расхождение незначительно и некритично. В целом обученная модель показывает хорошие результаты и обладает способностью к обобщению.

Результаты экспериментального исследования

Для проверки возможности снижения количества ложнопозитивных срабатываний проведен эксперимент, в котором сравнивалась работа классической архитектуры нейронной сети LSTM (Long Short-Term Memory) с добавлением модуля большой языковой модели (LLM). Коэффициент ложнопозитивных срабатываний (FP rate) вычислялся по формуле

$$\text{FP rate} = \frac{\text{FP}}{\text{FP} + \text{TN}}, \tag{10}$$

где FP – количество ложнопозитивных срабатываний; TN – количество верно классифицированных вредоносных пакетов.

Эксперимент проведен пять раз, и усредненные значения по пяти экспериментам представлены в таблице.

Результат сравнения
The result of the comparison

Модель	Входящие пакеты	Верно классифицированные пакеты	Ложнопозитивные срабатывания	Коэффициент ложнопозитивных срабатываний
LTSM	150 000	14 879	843	0,05361
LTSM + LLM	150 000	15 124	514	0,03286

Результаты показывают, что использование большой языковой модели в качестве дополнительного этапа проверки срабатываний системы обнаружения вторжений действительно позволяет снизить количество ложноположительных срабатываний на 38,64 %. Это значительное улучшение позволяет повысить надёжность системы обнаружения вторжений, уменьшая количество ложных тревог и, соответственно, снижая нагрузку на аналитиков безопасности. Проверено, может ли большая языковая модель улучшить работу системы обнаружения вторжений.

Заключение

В рамках данного исследования проведено экспериментальное исследование возможности использования больших языковых моделей (LLM) для снижения ложнопозитивных срабатываний в системах обнаружения вторжений (IDS). Важность данной темы продиктована стремительным ростом объема и сложности сетевых атак, что обусловлено экспоненциальным ростом сетевого трафика и данных в последние годы.

В ходе эксперимента построена и обучена базовая модель нейронной сети для быстрой классификации пакетов в сетевом трафике на вредоносные и безопасные. Особое внимание уделено архитектуре и работе LSTM-сетей, а также особенностям использования большой языковой модели Mixtral 8x7B, основанной на архитектуре разряженного экспертного состава. Для подготовки данных и обучения системы использовался набор данных CIC-IDS2017. Для улучшения обучения модели проведен отбор ключевых признаков из данной выборки. В ходе обучения LSTM-сети достигнуты высокие показатели точности и низкие потери, что свидетельствует об эффективности модели в классификации сетевых пакетов как вредоносных или безопасных.

Экспериментальные исследования подтвердили, что добавление модуля большой языковой модели позволяет значительно снизить количество ложнопозитивных срабатываний. Коэффициент ложнопозитивных срабатываний снизился на 38,64 %, что свидетельствует о существенном улучшении надежности системы обнаружения вторжений. Однако существенным недостатком является долгий ответ модели, который необходимо будет уменьшать в будущих исследованиях.

Список литературы/References

1. Odlyzko A.M. Internet Traffic Growth: Sources and Implications. *Proceedings of SPIE – The International Society for Optical Engineering*. 2003;5247. DOI: 10.1117/12.512942

2. Khraisat A., Gondal I., Vamplew P., Kamruzzaman J. Survey of intrusion detection systems: techniques, datasets and challenges. *Cybersecurity*. 2019;2:20. DOI: 10.1186/s42400-019-0038-7

3. Ho C.Y., Lin Y.R., Lai Y.C., Chen I.W., Wang F.Y., Tai W.H. False Positives and Negatives from Real Traffic with Intrusion Detection/Prevention Systems. *International Journal of Future Computer and Communication*. 2012;1(2):87–90. DOI: 10.7763/IJFCC.2012.V1.23

4. Naveed H., Khan A.U., Qiu S., Saqib M., Anwar S., Usman M., Akhtar N., Barnes N., Mian A. *A Comprehensive Overview of Large Language Models*. Preprint submitted to Elsevier, 2024.

5. Wan X., Liu H., Xu H., Zhang X. Network Traffic Prediction Based on LSTM and Transfer Learning. *IEEE Access*. 2022;10:86181–86193. DOI: 10.1109/ACCESS.2022.3199372

6. Xu H., Wang S., Li N., Wang K., Zhao Y., Chen K. et al. Large Language Models for Cyber Security: A Systematic Literature Review. arXiv preprint arXiv:2405.04760, 2024.

7. Markevych M., Dawson M. A Review of Enhancing Intrusion Detection Systems for Cybersecurity Using Artificial Intelligence (AI). *International Conference Knowledge-Based Organization*. 2023;29(3). DOI: 10.2478/kbo-2023-0072

8. Charalambous Y., Tihanyi N., Jain R., Sun Y., Ferrag M.A., Cordeiro L.C. A New Era in Software Security: Towards Self-Healing Software via Large Language Models and Formal Verification. arXiv:2305.14752 [cs.SE], 2023. DOI: 10.48550/arXiv.2305.14752

9. Alkhatib N., Mushtaq M., Ghauch H., Danger J.L. CAN-BERT do it? Controller Area Network Intrusion Detection System based on BERT Language Model. arXiv:2210.09439 [cs.LG], 2022. DOI: 10.48550/arXiv.2210.09439

10. Moskal S., Laney S., Hemberg E., O'Reilly U.M. LLMs Killed the Script Kiddie: How Agents Supported by Large Language Models Change the Landscape of Network Threat Testing. arXiv preprint arXiv:2310.06936v1 [cs.CR], 2023.

11. Temara S. Maximizing Penetration Testing Success with Effective Reconnaissance Techniques using ChatGPT. arXiv preprint arXiv:2307.06391 [cs.CR], 2023.

12. Pudjihartono N., Fadason T., Kempa-Liehr A.W., O'Sullivan J.M. A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. *Front Bioinform*. 2022;2:927312. DOI: 10.3389/fbinf.2022.927312

13. Xu S., Wu Z., Zhao H., Shu P., Liu Z., Liao W., Li S., Sikora A., Liu T., Li X. Reasoning Before Comparison: LLM-Enhanced Semantic Similarity Metrics for Domain Specialized Text Analysis. arXiv preprint arXiv:2402.11398v2 [cs.CL], 2024.

14. Van Houdt G., Mosquera C., Nápoles G. A Review on the Long Short-Term Memory Model. *Artificial Intelligence Review*. 2020;53(1). DOI: 10.1007/s10462-020-09838-1

15. Staudemeyer R.C., Morris E.R. Understanding LSTM – a tutorial into Long Short-Term Memory Recurrent Neural Networks. arXiv preprint arXiv:1909.09586v1, 2019.

16. Jiang A.Q., Sablayrolles A., Roux A., Mensch A., Savary B., Bamford C., Chaplot D.S., de las Casas D., Bou Hanna E., Bressand F., Lengyel G., Bour G., Lample G., Lavaud L.R., Saulnier L., Lachaux M.-A., Stock P., Subramanian S., Yang S., Antoniak S., Le Scao T., Gervet T., Lavril T., Wang T., Lacroix T., El Sayed W. Mixtral of Experts. arXiv preprint arXiv:2401.04088 [cs.LG], 2024.

17. Intrusion Detection Evaluation Dataset (CIC-IDS2017). Available at: https://www.unb.ca/cic/datasets/ids-2017.html/ (accessed 12.04.2023).

18. Cai J., Luo J., Wang S., Yang S. Feature selection in machine learning: A new perspective. *Neurocomputing*. 2018;300:70–79.

Информация об авторах

Болодурина Ирина Павловна, д-р техн. наук, проф., заведующий кафедрой прикладкой математики, Оренбургский государственный университет, Оренбург, Россия; ipbolodurina@yandex.ru.

Нефедов Дмитрий Алексеевич, аспирант кафедры прикладкой математики, Оренбургский государственный университет, Оренбург, Россия; namilaze@gmail.com.

Information about the authors

Irina P. Bolodurina, Dr. Sci. (Eng.), Prof., Head of the Department of Applied Mathematics, Orenburg State University, Orenburg, Russia; ipbolodurina@yandex.ru.

Dmitry A. Nefedov, Postgraduate student of the Department of Applied Mathematics, Orenburg State University, Orenburg, Russia; namilaze@gmail.com.

Вклад авторов: все авторы сделали эквивалентный вклад в подготовку публикации.

Авторы заявляют об отсутствии конфликта интересов.

Contribution of the authors: the authors contributed equally to this article.

The authors declare no conflicts of interests.

Статья поступила в редакцию 19.06.2024

The article was submitted 19.06.2024