МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»

Кафедра теоретических основ компьютерной безопасности и криптографии

Мультимодальные нейронные сети в системах информационной безопасности

АВТОРЕФЕРАТ
дипломной работы

студента 6 курса 631 группы
специальности 10.05.01 Компьютерная безопасность
факультета компьютерных наук и информационных технологий
Мызникова Сергея Анатольевича

Научный руководитель
Доцент _____ И. И. Слеповичев
19.01.2026 г.

Заведующий кафедрой
д. ф.-м. н., профессор _____ М. Б. Абросимов
19.01.2026 г.

Саратов 2026

ВВЕДЕНИЕ

Современные системы информационной безопасности функционируют в условиях постоянного усложнения и эволюции угроз. Кибератаки становятся более многоэтапными, адаптивными и могут затрагивать сразу несколько типов данных: сетевой трафик, программный код, веб-контент, мультимедийную информацию и другие. В таких условиях традиционные методы защиты, основанные на сигнатурах, правилах и анализе данных одной модальности, постепенно утрачивают эффективность.

Одной из наиболее серьёзных угроз информационной безопасности является веб-фишинг, который оказывает значительное влияние как на компании, так и на частных пользователей. Это вид атаки, при которой злоумышленник использует поддельные веб-страницы, имитирующие легитимные ресурсы, с целью кражи конфиденциальных данных пользователей: паролей, данных карт, учётных записей, личной информации. По данным портала Cybercrime Information Center, за период с мая 2024 года по апрель 2025 года было зарегистрировано около четырёх миллионов сообщений о фишинговой активности, среди которых было почти 1,5 миллиона уникальных доменов, подвергшихся фишингу.

При этом число фишинговых атак продолжает расти. Согласно данным портала Anti-Phishing Working Group, в первом квартале 2025 года было зарегистрировано порядка одного миллиона уникальных фишинговых сайтов, а во втором квартале показатель вырос до 1,13 миллиона.

Рост количества атак сопровождается усложнением их структуры и способов маскировки, в результате чего злоумышленники всё чаще комбинируют различные техники подмены контента, визуального обмана и манипуляции текстовой информацией.

Подход анализа, основанный на одном типе данных, ограничивает возможности выявления сложных и маскирующихся атак. В связи с этим всё большую актуальность приобретают методы, способные объединять разные модальности и выявлять взаимосвязи между ними. Одним из наиболее

перспективных направлений в данной области, которое развивается стремительными темпами, являются мультимодальные нейронные сети. Они позволяют в рамках единой модели совместно анализировать данные различных видов: текст, изображения, аудио и другие.

Мультимодальные нейронные сети способны выявлять несоответствия между модальностями и формировать более устойчивые и обобщённые признаки. Это делает такие модели особенно востребованными при решении задач информационной безопасности, где атаки часто проявляются неявно и в совокупности нескольких признаков.

Целью данной работы является исследование возможностей и оценка эффективности применения мультимодальных нейронных сетей в системах информационной безопасности. В рамках работы рассматриваются архитектурные особенности мультимодальных моделей, подходы к объединению различных модальностей и стратегии их практического использования для выявления угроз информационной безопасности.

Практическая часть работы заключается в оценке эффективности применения методов на основе мультимодальных нейронных сетей в задачах информационной безопасности на примере выявления фишинговых веб-страниц.

Дипломная работа состоит из введения, четырёх разделов, заключения, списка использованных источников и трёх приложений. Общий объем работы – 95 страниц, из них 55 страниц – основное содержание, включая две таблицы, список использованных источников из тридцати пяти наименований.

КРАТКОЕ СОДЕРЖАНИЕ

В разделе «Теория мультимодальных нейронных сетей» даётся описание понятия мультимодальности, приводятся механизмы работы мультимодальных нейронных сетей. Анализируются принципы обработки данных различных модальностей, методы их кодирования, выравнивания и объединения. Разобран подход к реализации, основанный на мультимодальных больших языковых моделях, который будет использоваться в практической части.

Исследуется роль кодировщиков модальностей, предназначенных для преобразования входных данных (изображений, аудио и других типов информации) в компактные векторные представления. Устанавливается, что на практике преимущественно используются предварительно обученные кодировщики, позволяющие сократить вычислительные затраты и повысить устойчивость моделей. Описаны подходы к увеличению разрешения входных данных, включая прямое масштабирование и разбиение на фрагменты.

Приводятся проблемы согласования признаков различных модальностей и методы их решения. Описывается использование специальных функций потерь, направленных на выравнивание эмбеддингов различных модальностей, в том числе рассматривается контрастивная функция потерь, позволяющая сближать семантически связанные данные и раздвигать несвязанные, формируя общее латентное пространство признаков.

Рассмотрены основные стратегии слияния: раннее, глубокое, позднее и гибридное. Приведены их преимущества и недостатки с точки зрения точности, вычислительной сложности и устойчивости к потере информации.

Описываются принципы работы больших языковых моделей на основе архитектуры трансформера, механизмы самовнимания, сети прямого распространения и методы нормализации. Предварительное обучение на больших текстовых корпусах позволяет большой языковой модели (БЯМ) обладать обобщёнными знаниями и способностью к рассуждению.

Рассмотрены методы интеграции нетекстовых модальностей в больших языковых моделях (БЯМ), описано два подхода к реализации интерфейса:

слияние на уровне токенов и слияние на уровне признаков. Также приводится альтернативный подход преобразования нетекстовых данных в текст с помощью специализированных моделей, имеющий преимущества простоты реализации и недостатки, связанные с потерей информации.

Рассмотрены этапы обучения мультимодальных больших языковых моделей (МБЯМ): предобучение, настройка под инструкции и выравнивание с предпочтениями пользователей. Изучен подход обучения и применения модели без подсказок и его значение для практического использования. Приведена формула авторегриссионной функции потерь, которая используется при обучении модели.

В разделе рассмотрены теоретические основы мультимодальных нейронных сетей и мультимодальных больших языковых моделей (МБЯМ). Показано, что такие модели позволяют эффективно объединять данные различных модальностей, обеспечивая более глубокий и устойчивый анализ.

В разделе «Проблемы информационной безопасности, решаемые мультимодальными нейронными сетями» рассматриваются основные угрозы информации, в решении которых может быть использован подход, основанный на мультимодальных нейронных сетях. Анализируются возможности мультимодальных подходов при выявлении фишинга, обнаружении подделок мультимедиа, детектировании вредоносного программного обеспечения, обнаружении сетевых вторжений и биометрической аутентификации. Раздел демонстрирует прикладную значимость мультимодальных моделей в современных системах защиты информации.

Исследуется задача выявления фишинга как одной из наиболее распространённых атак, основанных на методах социальной инженерии. Анализируются основные формы фишинговых атак и ограничения традиционных методов защиты из-за высокой изменчивости фишинговых ресурсов и возможности обхода.

Описана проблема обнаружения подделок мультимедиа, обусловленная развитием генеративных нейронных сетей и технологий создания

синтетического контента. Устанавливается, что унимодальные методы, ориентированные только на анализ изображения или аудио, оказываются недостаточно устойчивыми. Делается вывод, что мультимодальные нейронные сети, способные выявлять несоответствия между визуальными и аудиопризнаками, являются более надёжным инструментом для обнаружения подобных атак.

Рассмотрены подходы к анализу вредоносного ПО с использованием различных модальностей: байт-кода, изображений, графов вызовов. Показано, что вредоносные программы могут маскировать своё поведение, изменяя отдельные признаки, сохраняя при этом общую вредоносную логику. Обосновано применение мультимодальных нейронных сетей, которые позволяют объединять различные представления программ.

Приведены ограничения потоковых и пакетных систем обнаружения вторжений. Делается вывод о необходимости объединения различных источников информации, включая характеристики потоков и содержимое пакетов, с использованием мультимодальных нейронных сетей, что позволяет повысить полноту и точность обнаружения вторжений.

Рассматриваются недостатки одномодальных биометрических систем. Приводятся преимущества мультимодальных подходов в повышении точности и устойчивости к спуфингу.

В результате анализа делается вывод, что мультимодальные нейронные сети являются эффективным инструментом для решения широкого круга задач информационной безопасности. Их ключевым преимуществом является способность комплексного анализа разнородных данных и выявления скрытых взаимосвязей между модальностями, что обеспечивает более высокий уровень защиты по сравнению с традиционными одномодальными подходами.

В разделе «Веб-фишинг и методы противодействия» подробно рассматривается проблема веб-фишинга как объекта исследования. Анализируются принципы работы фишинговых веб-страниц, методы маскировки и воздействия на пользователя. Рассматриваются существующие

методы выявления фишинговых ресурсов, их преимущества и ограничения. Отдельное внимание уделяется мультимодальному анализу веб-страниц как наиболее устойчивому и перспективному подходу к обнаружению фишинга.

Рассмотрены основные механизмы фишинга: визуальный обман, манипуляции URL, подделка бренда, социальная инженерия и атаки в реальном времени.

Анализируются подходы, основанные на использовании чёрных списков, методах машинного обучения, исследовании структуры и содержимого HTML-кода, а также визуальном анализе веб-страниц. Устанавливается, что каждый из данных подходов обладает определёнными преимуществами, однако имеет и существенные ограничения. Чёрные списки не позволяют выявлять ранее неизвестные фишинговые ресурсы, методы машинного обучения чувствительны к изменению статистических признаков, анализ HTML-кода характеризуется высокой сложностью и неоднозначностью интерпретации, а визуальные методы требуют значительных вычислительных ресурсов и редко применяются изолированно.

Подчёркивается, что использование одномодальных методов делает системы обнаружения фишинга уязвимыми к обходу, поскольку злоумышленники могут целенаправленно изменять отдельные признаки веб-страниц. В результате эффективность таких систем снижается в условиях постоянного развития фишинговых методов.

Обосновывается целесообразность применения мультимодального анализа веб-страниц. Показано, что объединение анализа URL-адреса, HTML-кода и визуального представления страницы позволяет формировать более полное и устойчивое представление о веб-ресурсе. Мультимодальный подход обеспечивает выявление несоответствий между различными модальностями, снижает вероятность ложных срабатываний и повышает устойчивость системы к новым и модифицированным видам фишинговых атак.

В результате делается вывод, что мультимодальный анализ является наиболее перспективным направлением для построения современных систем

обнаружения веб-фишинга, способных эффективно функционировать в условиях высокой динамики угроз и разнообразия методов маскировки, используемых злоумышленниками.

Раздел «Практическая часть» посвящён практической реализации и экспериментальной оценке системы выявления фишинговых веб-страниц на основе мультимодальных больших языковых моделей. Описываются используемые модели, организация датасета, процесс подготовки входных данных и формирование промптов. Проводится тестирование моделей, анализируются метрики качества и реализуется API для автоматизированной проверки веб-страниц.

В начале раздела обосновывается выбор мультимодальных больших языковых моделей в качестве инструмента анализа веб-страниц. Отмечается, что в отличие от классических мультимодальных методов машинного обучения, требующих явного извлечения признаков и дополнительного обучения, мультимодальные большие языковые модели (МБЯМ) позволяют выполнять совместный анализ различных модальностей в режиме без подсказок. Это делает подход менее зависимым от специализированных датасетов и более устойчивым к появлению новых типов фишинговых атак.

Рассматриваются используемые мультимодальные модели, состоящие из qwen3-vl, llava, gemma3, mistral-small3.1, и их архитектурные особенности. Приводятся параметры запуска моделей, направленные на получение детерминированных результатов и обеспечение сопоставимости экспериментов.

Описывается организация экспериментального датасета, включающего фишинговые и легитимные веб-страницы, а также структура хранения данных, объединяющая URL-адрес, HTML-код страницы и её визуальное представление в виде скриншота.

Описывается процесс взаимодействия с мультимодальной модель и механизм передачи текстовых данных и изображений в модель, а также использование промптов для формализации задачи классификации. Логика выявления фишинга задаётся не путём изменения архитектуры модели или её

дообучения, а посредством чётко сформулированной инструкции, определяющей допустимые источники информации, критерии принятия решений и формат выходных данных.

Показано, что классификация веб-страниц осуществляется в режиме без подсказок, при котором модель не обучается дополнительно на специализированных фишинговых датасетах. Она опирается на обобщённые знания, полученные в ходе предварительного обучения, и выполняет анализ страницы на основе смысловых, структурных и визуальных несоответствий. Такой подход снижает зависимость от актуальности обучающих данных и повышает устойчивость метода к появлению новых и модифицированных видов фишинговых атак.

В рамках экспериментальной части проводится тестирование моделей на задаче выявления фишинговых веб-страниц и задаче определения имитации или подделки бренда. Описываются показатели истинно положительных, истинно отрицательных, ложно положительных и ложно отрицательных срабатываний, на основе которых рассчитываются стандартные метрики точности, полноты и F1-меры.

Рассматриваются результаты тестирования мультимодальных моделей на задаче выявления фишинговых веб-страниц. Анализируются полученные значения метрик и выявляются особенности поведения различных моделей.

Отдельно приведены результаты анализа работы моделей в задаче выявления имитации или подделки бренда, которая рассматривается как частный случай фишинга. Данная задача требует более точного сопоставления информации между различными модальностями, поскольку название бренда, визуальная идентичность и доменное имя могут отличаться по форме представления. При переходе от общей задачи выявления фишинга к задаче определения подделки бренда поведение моделей изменяется, что отражается в перераспределении значений точности и полноты.

На основе сравнительного анализа результатов делается вывод о том, что модель Mistral-small3.1 демонстрирует наиболее сбалансированные показатели

по всем основным метрикам как в задаче выявления фишинга, так и в задаче определения имитации бренда. Высокие значения F1-меры свидетельствуют о способности модели эффективно обнаруживать фишинговые веб-страницы без существенного увеличения числа ложных срабатываний.

Проведённое тестирование подтверждает возможность практического применения мультимодальных больших языковых моделей для задач информационной безопасности. Полученные результаты демонстрируют, что использование мультимодального анализа в режиме без подсказок позволяет достигать высокого качества классификации без необходимости специализированного обучения и ручного проектирования признаков.

В заключительной части раздела описывается реализация программного интерфейса приложения для автоматизированной проверки веб-страниц. Рассматривается полный цикл работы системы, включающий загрузку HTML-кода страницы, формирование визуального представления, подготовку входных данных и передачу их в мультимодальную нейронную сеть. Показано, что разработанный программный интерфейс позволяет автоматизировать процесс анализа и использовать предложенный подход в системах информационной безопасности.

ЗАКЛЮЧЕНИЕ

В рамках данной дипломной работы были исследованы возможности применения мультимодальных нейронных сетей в системах информационной безопасности. В работе было показано, что подходы, основанные на анализе данных одной модальности, обладают ограниченной эффективностью при выявлении сложных атак, использующих совокупность различных признаков.

В ходе выполнения работы были рассмотрены архитектурные особенности мультимодальных нейронных сетей, а также подходы к объединению разнородных типов данных, включая текстовую и визуальную информацию.

В практической части работы было проведено сравнение нескольких локальных мультимодальных моделей в задачах выявления фишинговых веб-страниц. Для тестирования был выбран датасет, включающий как фишинговые, так и легитимные сайты, а в качестве входных данных использовались URL-адрес, HTML-код страницы и её визуальное представление в виде скриншота.

Результаты показали, что использование мультимодального подхода позволяет эффективно выявлять фишинговые веб-страницы. Наиболее сбалансированные результаты продемонстрировала модель Mistral-small3.1, показав высокие значения метрик при умеренных требованиях к вычислительным ресурсам.

Помимо этого, в рамках работы был реализован API для автоматической проверки страницы при помощи мультимодальных нейронных сетей, по полученному URL-адресу. Что позволяет пользователю выполнить быструю проверку страницы без необходимости собирать данные вручную.

Результаты работы подтверждают эффективность использования мультимодальных нейронных сетей в задачах по выявлению фишинговых веб-страниц на основе визуальной и текстовой информации. Такой подход позволяет с высокой вероятностью определять как частные, так и общие случаи фишинговых атак. Полученные результаты можно будет использовать для совершенствования систем информационной безопасности.