АТАКИ И МЕТОДЫ ЗАЩИТЫ ПРИ ИСПОЛЬЗОВАНИИ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ В КОНТЕКСТЕ СТЕГОАНАЛИЗА ЦИФРОВОГО КОНТЕНТА

Д.И. Сивков, *Национальный исследовательский университет ИТМО, sivkov@itmo.ru;*
М.Ю. Федосенко, *Национальный исследовательский университет ИТМО, аспирант, fedosenkomaksim98@gmail.com.*

УДК 004.056

Аннотация. Статья освещает роль методов машинного обучения в стегоанализе – выявлении скрытой информации в цифровых носителях. Описывается, как различные методы, в том числе глубокие нейронные сети и графовые сети, используются для улучшения процессов обнаружения стеганографии, обеспечивая более эффективную защиту данных. Рассматривается их применение, преимущества и ограничения, а также риски, связанные с атаками на машинное обучение и методы защиты.

Ключевые слова: стеганография; машинное обучение; нейронные сети; атаки отравления; атаки с уклонением; атаки на модель.

ATTACKS AND DEFENSE METHODS WHEN USING MACHINE LEARNING METHODS IN THE CONTEXT OF STEGOANALYSIS OF DIGITAL CONTENT

D. Sivkov, *National Research University ITMO;*
M. Fedosenko, *National Research University ITMO.*

Annotation. The article highlights the role of machine learning methods in steganalysis – identifying hidden information in digital media. Describes how various techniques, including deep neural networks and graph networks, are used to improve steganography detection processes, providing more effective data protection. Reviews their applications, advantages and limitations, and risks associated with machine learning attacks, and discusses defense methods.

Keywords*:* steganography; machine learning; neural networks; poisoning attacks; evasion attacks; attacks on the model.

Введение

В нашем быстро меняющемся цифровом мире стеганография и стегоанализ становятся все более важными понятиями в области информационной безопасности. В то время как стеганография предоставляет способы для скрытой передачи данных, стегоанализ стремится эти способы обнаружить и анализировать. С развитием машинного обучения возникают новые подходы к распознаванию и классификации скрытых сообщений, что позволяет раскрыть новые горизонты в обнаружении цифровых угроз и обеспечении конфиденциальной информации [1].

Целью данной работы является анализ методов машинного обучения на применимости в задачах стеганоанализа. Каждый из данных методов имеет свои особенности и сферы применения, что подчеркивает необходимость тщательного выбора инструментов в зависимости от конкретных задач и типов данных. Работа содержит в себе решение следующих задач:

1. Анализ особенностей методов машинного обучения с выделением их достоинств и недостатков для задач обработки стеганоконтейнеров.

2. Выявление подходов искусственного интеллекта (*AI*) и соответствующих им технологий под конкретные виды контейнеров: изображение, текст, файлы.
3. Разбор возможных атак на системы искусственного интеллекта применительно к интеллектуальным системам стеганоанализа.
4. Формулирование подходов защиты от атак на интеллектуальные системы стеганоанализа со стороны вредоносных изменений обрабатываемого информационного контента.

Таблицы, представленные в данной статье, демонстрируют сравнительный анализ разнообразных методов машинного обучения, каждый из которых привносит свой вклад в понимание применения методов машинного обучения в стеганографии, учитывая при этом свои уникальные особенности и области применения. Эти методы варьируются от глубоких нейронных сетей до графовых нейронных сетей, отражая многообразие подходов и спецификаций, необходимых для успешного стегоанализа. Однако не стоит забывать о присущих им ограничениях и сложностях, которые могут возникнуть при работе со сложными задачами обнаружения. Вместе с возможностями, методы машинного обучения приносят и угрозы, такие как атаки с уклонением, отравления данных и модельные атаки, которые могут подорвать эффективность стегоанализа в контексте разработки интеллектуальной модели анализа контента. В ответ на эти вызовы, применяются различные методы защиты, исследование которых также является целью данной работы.

Теоретические основы машинного обучения в стегоанализе

Стегоанализ, сложный процесс анализа скрытой информации в цифровых носителях, в последнее время активно развивается благодаря применению методов машинного обучения [2]. Важность теоретической базы в данной области сложно переоценить, так как она лежит в основе понимания, каким образом алгоритмы могут «учиться» распознавать и анализировать стеганографические вложения.

Применение различных методов машинного обучения в стегоанализе открывает новые перспективы для распознавания и классификации скрытых сообщений в цифровых средствах. Табл. 1 представляет собой сравнение нескольких подходов, каждый из которых имеет свои уникальные особенности и области применения.

Таблица 1.

Методы машинного обучения	**Подходит для**	**Особенности**
Глубокие нейронные сети [3]	Общего анализа	Многослойная обработка
Сверточные нейронные сети [4]	Изображений и видео	Специализирован на визуальных данных
Рекуррентные нейронные сети [5]	Последовательностей (текст, аудио)	Учитывает временную последовательность
Сети долгой краткосрочной памяти [6]	Длительных последовательностей	Эффективен для долгосрочных зависимостей

Методы машинного обучения	Подходит для	Особенности
Автокодировщики [7]	Реконструкции и обнаружения	Использует обучение без учителя
Ограниченные машины Больцмана [8]	Вероятностного моделирования	Эффективен для извлечения признаков
Глубокая сеть доверия [9]	Сложных вероятностных моделей	Мощный, но сложный в настройке
Графовые нейронные сети [10]	Анализа графов и сетей	Подходит для структурированных данных

Из представленного сравнения видно, что каждый метод машинного обучения имеет свои сильные стороны в зависимости от контекста применения. Выбор метода зависит от множества факторов, включая тип и структуру данных, требуемую точность и сложность моделирования. Сбалансированный выбор подхода к машинному обучению позволяет повысить эффективность процесса стегоанализа, одновременно оптимизируя ресурсы и время обработки.

Однако, как и любая технология, методы машинного обучения не лишены недостатков, особенно когда речь идет о сложных задачах стегоанализа. В табл. 2 приведены ключевые недостатки различных методов машинного обучения в стегоанализе.

Таблица 2.

Методы машинного обучения	Недостатки
Глубокие нейронные сети [3]	Требуют большого количества данных для обучения и сложны в настройке
Сверточные нейронные сети [4]	Не эффективны для обработки временных последовательностей и анализа неструктурированных данных
Рекуррентные нейронные сети [5]	Могут страдать от проблемы затухания градиента и сложны в работе с очень длинными последовательностями
Сети долгой краткосрочной памяти [6]	Сложны в настройке и тренировке, а также требовательны к вычислительным ресурсам
Автокодировщики [7]	Могут неэффективно работать с сильно зашумленными данными и риск переобучения
Ограниченные машины Больцмана [8]	Сложны в понимании и настройке, а также могут быть неэффективными для очень сложных задач

Методы машинного обучения	Недостатки
Глубокая сеть доверия [9]	Требуют значительных вычислительных ресурсов и сложны в настройке, а также интерпретации
Графовые нейронные сети [10]	Ограничены в обработке неструктурированных данных, а также сложность в интерпретации результатов

Анализ недостатков, представленных в табл. 2, подчеркивает значимость тщательного планирования и тестирования при внедрении методов машинного обучения в практику стегоанализа. Понимание и принятие во внимание ограничений каждого метода является критически важным для минимизации ошибок и увеличения надежности результатов. Это также подчеркивает важность постоянного развития и адаптации моделей машинного обучения в соответствии с постоянно меняющимися требованиями информационной безопасности.

Методы машинного обучения для стегоанализа цифрового контента

На основе анализа случаев использования технологий искусственного интеллекта в задачах стеганоанализа выделены технологии, применение которых целесообразно как для несбалансированного характера набора данных контента интернет-ресурсов, так и для специфики данной отрасли. Конкретные технологии для подходов *Artificial Intelligence* в задачах стеганоанализа [11] в зависимости от покрывающего объекта представлены в табл. 3.

Таблица 3.

Вид вложения	Подход *AI*	Технология
Изображение	Методы обработки изображений (компьютерное и машинное зрение)	Рендеринг Сопоставление шаблонов Бинаризация Сегментация Счетчик пикселей Обнаружение и измерение краев изображений Оптическое распознавание Анализ признаков (текстуры, гистограммы) Пространственная и фильтрация Фильтры Габора Фильтры *ICA* Детекторы *LoG*, *DoG*, Харриса Дескриптор *SIFT* частотная
Лингвистическая стеганография	Обработка естественного языка (*Natural language processing*)	*AlchemyAPI*, *Expert System S.p.A.*, *General Architecture for Text Engineering* (*GATE*), *Modular Audio Recognition Framework, MontyLingua*, *Natural Language Toolkit* (*NLTK*)

Вид вложения	Подход *AI*	Технология
Цифровая стеганография	Методы машинного обучения (с учителем, без учителя)	Наивный Байес, деревья решений, логистическая регрессия, *k*-ближайших соседей, метод опорных векторов, линейная регрессия, полиномиальная регрессия, Метод *k*-средних, *Mean-Shift, DBSCAN, SOM*
Новые виды сокрытия информации	Нейронные сети (свёрточные сети)	Нейронная сеть Хопфилда Самоорганизующаяся карта Кохонена Нейронная сеть Ворда Сеть Хэмминга Сеть Элмана Многослойный перцептрон Перцептрон Розенблатта Когнитрон *AlexNet,* *ResNets*, *VGGs*, *Inception*

Исходя из анализа технологий, можно определить следующие области применения искусственного интеллекта, которые особенно подходят для стеганоанализа:

- Классификация: этот процесс включает в себя отнесение объектов к определенным классам на основании их атрибутов, где классы известны заранее. Эта задача решается с использованием методов обучения с учителем.
- Кластеризация: задача заключается в выявлении групп объектов (кластеров) по их характеристикам, при этом количество и свойства кластеров не заданы заранее. Решается с помощью методов обучения без учителя и нейронных сетей.
- Регрессия: предполагает определение значений атрибутов объекта на основе его взаимосвязей с другими объектами. Методы как обучения с учителем, так и без учителя используются для выявления этих зависимостей.
- Прогнозирование: включает как классификацию объектов и регрессионный анализ их характеристик, так и определение конечных объектов и их свойств. Методы машинного обучения применяются для первой категории задач, а нейронные сети – для второй.
- Аппроксимация: ищет функцию, которая максимально точно описывает данные или упрощает их для анализа взаимосвязей и создания новых объектов. Применяются нейронные сети для моделирования процессов мышления и принятия решений на основе ограниченных данных.
- Сжатие данных: направлено на устранение избыточности данных для экономии ресурсов и улучшения точности прогнозов ИИ, используя алгоритмы минимизации ошибок. Задача решается через ассоциативную память и алгоритмы обратного распространения ошибки в нейронных сетях, а также через предварительную обработку данных.
- Оптимизация: заключается в выборе наилучших параметров для создания эффективной и экономичной модели ИИ. Решается через предварительную обработку данных и использование самоорганизующихся нейронных сетей для оптимизации конкретных моделей ИИ [11].

Безопасность методов машинного обучения применительно к стегоанализу

Модели машинного обучения, хотя и являются мощными инструментами для обнаружения стеганографических сообщений, также могут быть уязвимы для ряда атак. Угрозы безопасности могут подорвать целостность и надежность систем стегоанализа, делая защиту моделей критически важной задачей.

Типы угроз и атак на модели машинного обучения

1. Атаки с уклонением представляют собой методы, используемые для обмана систем обнаружения стеганографии, обычно реализуемых с помощью алгоритмов машинного обучения. Цель таких атак – модифицировать стеганографический контент таким образом, чтобы он не был обнаружен как подозрительный [12].

Виды атак с уклонением:

1) Модификация изображений, предположим, есть изображение, в которое встроено стеганографическое сообщение. Стеганоанализатор, обученный на определенном наборе признаков, может обнаружить такие изменения. Атакующий может добавить в изображение специальный шум или применить техники обработки изображений, чтобы изменить эти признаки и сделать их менее заметными для системы стегоанализа, сохраняя при этом встроенное сообщение.
2) Адаптивные методы стеганографии изменяют способ встраивания сообщения в зависимости от содержимого носителя. Например, сообщение может быть встроено в области изображения с высокой текстурой, где изменения менее заметны. Таким образом, даже если стеганоанализатор обучен распознавать изменения в низкотекстурных областях, атака с уклонением будет направлена на уклонение от этих областей, уменьшая вероятность обнаружения.
3) Генеративно-состязательные сети могут быть использованы для создания стеганографических изображений, которые выглядят естественно и не вызывают подозрений у систем машинного обучения. *GAN* обучается генерировать изображения, которые не только содержат стеганографическую информацию, но и успешно обходят обнаружение путем имитации нормального распределения признаков в немодифицированных изображениях.

2. Атаки отравления включают намеренное введение некорректных, манипулированных или вредоносных данных в обучающий набор данных, с которым работает система обнаружения стеганографии. Целью такой атаки является искажение процесса обучения модели машинного обучения, чтобы ослабить ее способность правильно классифицировать данные и обнаруживать стеганографические сообщения в будущем [13].

Разновидности атак с отравлением данных:

1) Атакующий может предоставить исследователям или системам безопасности стеганографические изображения, которые ошибочно помечены как чистые. Если эти изображения используются в обучающем наборе данных, то обученная на них модель может стать менее чувствительной к реальным стеганографическим вложениям.
2) Злоумышленник может внедрить в обучающий набор сложноструктурированные образцы, которые маскируются под безопасный контент, но при этом содержат стеганографические данные. Таким образом,

модель обучается рассматривать подобные шаблоны как нормальные, что снижает точность будущего обнаружения.

3. Атаки на модель представляют собой вид атак на системы машинного обучения, где атакующий использует выходные данные модели для восстановления информации о входных данных или о самой модели. Эти атаки могут быть использованы для восстановления стеганографического сообщения, скрытого в носителе, или для получения информации о параметрах стеганографического алгоритма [14].

Разновидности атак на модели машинного обучения:

1) Если атакующий имеет доступ к стегоанализатору, который определяет наличие стеганографического сообщения в носителе, он может попытаться изменять входные данные и наблюдать за реакцией системы, чтобы восстановить скрытое сообщение. Например, изменяя пиксели изображения и изучая изменения в выходных данных модели, атакующий может вывести характеристики встроенного сообщения и даже восстановить его содержимое.
2) Атакующий может использовать доступ к модели стегоанализа, чтобы определить специфические характеристики стеганографического алгоритма, используемого для скрытия информации. Путем постепенного модифицирования входных данных и анализа выходных данных, можно выявить, какие параметры алгоритма были использованы для встраивания сообщения, что позволит эффективнее создавать стеганографический контент, который трудно обнаружить.
3) В ситуации, когда атакующий пытается обнаружить, как модель делает прогнозы относительно наличия или отсутствия стеганографии, он может использовать различные входные данные для создания «карты» решений модели. Это может включать в себя попытку определить границы решения или даже восстановить функцию потерь, используемую при обучении модели.

Способы защиты от атак

1. Состязательное обучение – это метод обучения машинного обучения, направленный на повышение устойчивости моделей к атакам с уклонением, в том числе к состязательным примерам. Этот метод включает в обучающий процесс примеры, специально разработанные для введения модели в заблуждение, тем самым заставляя ее «учиться» на своих ошибках и становиться более устойчивой к подобного рода атакам [15].

Способы применения состязательного обучения:

1) Исследователь может генерировать изображения, в которые встроены стеганографические сообщения с использованием различных алгоритмов и техник. Затем эти изображения могут быть искусственно модифицированы для создания состязательных примеров, которые затрудняют обнаружение встроенных сообщений существующими методами стегоанализа. Включая эти модифицированные изображения в набор данных для обучения, можно улучшить способность модели обнаруживать стеганографию, даже если она была специально адаптирована для уклонения от детекции.
2) Создается модель машинного обучения, которая обучается обнаруживать стеганографические сообщения в изображениях. В процессе обучения модели предъявляются не только обычные примеры стеганографии, но и адверсариальные примеры, созданные для обхода детекции. Это позволяет

модели адаптироваться к разнообразным стратегиям скрытия информации и повышает ее эффективность в условиях реального использования.

3) После обучения модель стегоанализа может быть протестирована на новом наборе состязательных примеров, чтобы оценить ее устойчивость к атакам с уклонением. Этот подход позволяет идентифицировать потенциальные слабости в модели и дополнительно улучшить ее защищенность.

2. Регуляризация в машинном обучении – это метод, направленный на предотвращение переобучения модели за счет добавления дополнительного ограничения (штрафа) на величину весов модели. Это помогает улучшить обобщающую способность модели на новых, невиданных данных, делая ее менее чувствительной к шуму в обучающем наборе данных. В контексте стеганографии и стегоанализа регуляризация может использоваться для улучшения устойчивости и точности моделей, обнаруживающих стеганографические вложения [16].

Способы применения регуляризации:

1) При разработке модели машинного обучения для обнаружения стеганографических сообщений в мультимедийных файлах, таких как изображения или аудио, использование регуляризации помогает предотвратить переобучение на особенности конкретного набора данных. Например, *L*1-регуляризация (*Lasso*) может быть использована для обеспечения разреженности весов модели, что полезно для выявления наиболее значимых признаков, указывающих на наличие стеганографии, в то время как *L*2-регуляризация (*Ridge*) помогает снизить общую величину весов, делая модель менее чувствительной к шуму в данных.
2) В сценариях, где генеративные модели, такие как генеративно-состязательные сети, используются для создания стеганографических вложений, регуляризация может способствовать генерации более естественно выглядящих изображений или аудиофайлов, в которые встроено скрытое сообщение. Это уменьшает вероятность обнаружения встроенной информации аналитическими инструментами.
3) Разработчики стеганографических алгоритмов могут применять регуляризацию для минимизации изменений, вносимых в контейнер (например, в изображение), тем самым снижая обнаружимость стеганографического сообщения. Например, методы, основанные на минимизации общего вариационного регуляризатора, могут использоваться для сохранения гладкости и структурных особенностей изображения при встраивании в него информации, делая модификации менее заметными для стегоанализа.

3. Ансамблевые методы в машинном обучении – это подходы, при которых для принятия окончательного решения используются несколько обучающих моделей. Основная идея состоит в том, что комбинация предсказаний от множества моделей приведет к лучшей производительности, чем использование любой отдельной модели. Это достигается за счет уменьшения дисперсии (*variance*), смещения (*bias*) и ошибок из-за случайных флуктуаций в обучающем наборе данных [17].

Кейсы применения ансамблевых методов:

1) В контексте стегоанализа ансамблевые методы могут использоваться для комбинирования различных детекторов стеганографии. Например, можно использовать ансамбль из разных типов нейронных сетей, таких как сверточные нейронные сети (*CNN*) для обработки изображений и рекуррентные нейронные сети (*RNN*) для обработки временных

последовательностей аудиофайлов, чтобы повысить точность обнаружения стеганографических сообщений.

2) Иногда одна модель лучше работает с определенными типами стеганографии, в то время как другая модель лучше справляется с другими типами. Создание ансамбля, включающего различные методы стегоанализа, такие как анализ наименьших значащих битов (*LSB analysis*) и частотный анализ, может обеспечить более всестороннее обнаружение.
3) Методы усиления, такие как *AdaBoost*, могут использоваться для комбинирования нескольких слабых классификаторов стеганографии в более мощный ансамбль. Например, несколько простых детекторов, каждый из которых способен обнаружить только определенные виды стеганографических вмешательств, могут быть объединены, чтобы создать систему, способную эффективно обнаруживать широкий спектр стеганографических методов.

Заключение

Стегоанализ представляет собой динамично развивающуюся область, в которой применение методов машинного обучения в будущем будет играть важную роль. В результате анализа интеллектуальных подходов, таблицы 1 и 2 показывают, что каждый из методов обучения обладает определенными преимуществами и недостатками, варьирующимися в зависимости от типа данных и специфики задачи. Важно осознавать, что ни один метод не является универсальным решением, и именно сочетание различных подходов часто приводит к наиболее эффективным системам стегоанализа.

С другой стороны, необходимо учитывать потенциальные угрозы и уязвимости, связанные с применением машинного обучения в области стеганографии. Учет этих угроз крайне важен при разработке комплексной программной системы интеллектуального стеганоанализа информационного контента, поскольку злоумышленник за счет изменения данных способен нарушить процесс работы данного компонента защиты [11]. Осведомленность об атаках с уклонением, отравления данных и атаках на модель позволяет разработать защитные механизмы, такие как состязательное обучение, регуляризация и ансамблевые методы, которые повышают устойчивость и надежность моделей. Эти методы позволяют не только улучшить точность обнаружения, но и обеспечить защиту от современных и сложных форм атак. Следует также отметить, что применение состязательного обучения, регуляризации и ансамблевых методов может существенно повысить способность моделей выявлять и противостоять новым видам и формам стеганографии, обеспечивая тем самым более надежную защиту информации.

В перспективе, сочетание продвинутых методов машинного обучения с глубоким пониманием специфики стегоанализа открывает новые возможности для создания более эффективных и устойчивых систем защиты информации, вносит вклад в такие научные отрасли, как криптографическая защита информации и искусственный интеллект, реализует приоритет стратегии научно-технологического развития РФ «Переход к передовым технологиям проектирования и создания высокотехнологичной продукции, основанным на применении интеллектуальных производственных решений, роботизированных и высокопроизводительных вычислительных систем, новых материалов и химических соединений, результатов обработки больших объемов данных, технологий машинного обучения и искусственного интеллекта» [18].

Литература

1. Юренский П.В. Методы статистического и нейросетевого стегоанализа скрытых каналов // Инновации в науке, 2019. – № 1. – С. 11-13.
2. Дрюченко М.А., Сирота А.А. Стегоанализ цифровых изображений с использованием глубоких нейронных сетей и гетероассоциативных интегральных преобразований // Прикладная дискретная математика, 2022. – № 55. – С. 36-56.
3. Созыкин А.В. Обзор методов обучения глубоких нейронных сетей // Вестник ЮУрГУ. Серия: Вычислительная математика и информатика, 2017. – № 3. – С. 28-59.
4. Скрипачев В. О., Гуйда М. В., Гуйда Н. В., Жуков А. О. Особенности работы сверточных нейронных сетей // International Journal of Open Information Technologies, 2022. – № 12. – С. 53-60.
5. Андросова Е.Е. Применение рекурсивных рекуррентных нейронных сетей // Новые информационные технологии в автоматизированных системах, 2016. – № 19. – С. 107-114.
6. Пустынный Я.Н. Решение проблемы исчезающего градиента с помощью нейронных сетей долгой краткосрочной памяти // Инновации и инвестиции, 2020. – № 2. – С. 130-132.
7. Ваняшкин Ю.Ю., Макаров Д.А., Попова И.А., Соболева Е.Д. Применение автокодировщиков для устранения шумов с изображений // StudNet, 2020. – № 10. – С. 27-38.
8. Абросимов М.А., Бровко А.В. Метод обучения слоев свертки в искусственной нейронной сети с помощью ограниченной машины Больцмана // Вестник СГТУ, 2015. – № 1. – С. 114-117.
9. Татьянкин В.М., Дюбко И.С. Нейронные сети глубокого доверия в сравнение с многослойным персептроном // Вестник ЮГУ, 2015. – № 2. – С. 87-89.
10. Циликов Н.С., Федосин С.А. Графовые нейронные сети // Вестник МГУ, 2012. – № 2. – С. 161-163.
11. Федосенко М.Ю., Беззатеев С.В. Анализ проблематики применения методов стеганографии при осуществлении противоправных действий и ее роли в цифровой криминалистике // Проблемы информационной безопасности. Компьютерные системы, 2023. – № 3. – С. 33-57.
12. Костюмов В.В. Обзор и систематизация атак уклонением на модели компьютерного зрения // International Journal of Open Information Technologies, 2022. – № 10. – С. 11-20
13. Намиот Д.Е. Введение в атаки отравлением на модели машинного обучения // International Journal of Open Information Technologies, 2023. – № 3. – С. 58-68.
14. Намиот Д.Е. Схемы атак на модели машинного обучения // International Journal of Open Information Technologies, 2023. – № 5. – С. 68-86.
15. Li H., Namiot D. A Survey of adversarial attacks and defenses for image data on deep learning // International Journal of Open Information Technologies, 2022. – № 5. – С. 9-16.
16. Тимофеева О.П., Неимущев С.А., Неимущева Л.И., Тихонов И.А. Распознавание эмоций по изображению лица на основе глубоких нейронных сетей // Труды НГТУ им. Р. Е. Алексеева, 2020. – № 1. – С. 16-24.
17. Фирюлина М.А., Каширина И.Л. Описание процесса прогнозирования проблемных состояний с применением ансамблевых методов машинного обучения // Инженерный вестник Дона, 2022. – № 4. – С. 34-46.
18. URL https://нтр.рф/challenges-priorities/ (дата обращения – июль 2024 г.).