МИНОБРНАУКИ РОССИИ

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

Кафедра информатики и программирования

ГЕНЕРАТИВНО-СОСТЯЗАТЕЛЬНЫЕНЕЙРОННЫЕ СЕТИ В СТЕГАНОГРАФИИ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 441 группы
направления 02.03.03 Математическое обеспечение и администрирование информационных систем
факультета компьютерных наук и информационных технологий
Козикова Данилы Алексеевича

Научный руководитель:
доцент _____ Кудрина Е.В.

Зав. кафедрой:
к.ф-м.н, доцент _____ Огнёва М.В.

Саратов 2026

ВВЕДЕНИЕ

Современное развитие цифровых коммуникаций сопровождается резким увеличением объёмов передаваемой информации, что неизбежно приводит к росту угроз её несанкционированного доступа и анализа. В этих условиях особое значение приобретают методы скрытой передачи данных, обеспечивающие не только защиту содержимого, но и сам факт существования сообщения. Данная задача решается средствами стеганографии, основной целью которой является внедрение информации в цифровые носители: изображения, аудио- и видеофайлы – без заметного изменения их структуры.

Классические подходы к стеганографии, такие как методы LSB (Least Significant Bit) в пространственной области или DCT/DWT в частотной, отличаются простотой реализации, но характеризуются низкой устойчивостью к внешним воздействиям – сжатию, фильтрации и добавлению шума. Даже незначительные изменения носителя могут привести к утрате внедрённого сообщения. Более современные адаптивные методы, учитывающие локальные статистические свойства изображения (например, HUG или WOW), позволяют повысить устойчивость, однако требуют ручной настройки параметров и не всегда обеспечивают оптимальный баланс между незаметностью, ёмкостью и устойчивостью.

С развитием глубинного обучения и появлением генеративно-состязательных нейронных сетей (Generative Adversarial Networks, GAN) возникли новые возможности для решения задач стеганографии. GAN способны обучаться формировать изображения с внедрённой информацией таким образом, что они практически неотличимы от оригинальных не только визуально, но и для специализированных систем стегоанализа. Это открывает перспективы создания интеллектуальных стеганографических систем, автоматически подстраивающихся под структуру изображений и характеристики атак.

Таким образом, разработка стеганографической системы на основе

генеративно-состязательных нейронных сетей является актуальной научно-практической задачей, находящейся на пересечении направлений информационной безопасности, компьютерного зрения и искусственного интеллекта.

Выше сказанное определило цель работы – разработать стеганографическую систему, использующую генеративно-состязательную нейронную сеть, для внедрения и извлечения скрытых сообщений в изображениях, а также провести исследование эффективности данной системы на основе стеганографических метрик.

Поставленная цель определила следующие задачи:

1. Проанализировать классические и современные методы стеганографии.
2. Изучить архитектуру и принципы обучения генеративно-состязательных сетей.
3. Рассмотреть инструментальные системы и технологии, применяемые при разработке нейронных стеганографических систем.
4. Спроектировать и реализовать нейросетевую стеганографическую систему для внедрения и извлечения информации в изображениях.
5. Продемонстрировать работу нейросетевой стеганографической системы и оценить её эффективность.

Методологическую основу исследования составляют труды в области цифровой стеганографии и нейросетевых методов защиты информации: А.В. Петрухина, S.N.M. Al-Faydi, S.K. Ahmed, S. Aghili, A. Jahangir, V. Holub, J. Fridrich, В.В. Побыванца, Е.Г. Барщевского.

Теоретическая значимость работы состоит в систематизации и сравнительном анализе классических и нейросетевых подходов к стеганографии изображений, а также в обосновании применения архитектуры Encoder–Decoder–Discriminator на базе GAN для задачи скрытой передачи данных.

Практическая значимость заключается в разработке и программной реализации стеганографической системы SteganoGAN на базе PyTorch,

достигшей значений PSNR = 41,84 дБ, SSIM = 0,996 и точности декодирования 99,7%, что подтверждает применимость системы в задачах защиты авторских прав и конфиденциальной передачи данных.

Структура и объём работы. Бакалаврская работа состоит из введения, двух разделов, заключения, списка использованных источников (17 наименований) и 11 приложений. Общий объём работы – 97 страниц, основное содержание включает 8 рисунков.

КРАТКОЕСОДЕРЖАНИЕ РАБОТЫ

В первом разделе «Теоретическая часть» осуществлен комплексный теоретический анализ предметной области, проведена понятийно-математическая формализация процессов скрытия данных и исследованы фундаментальные принципы построения генеративно-состязательных нейронных сетей для задач стеганографии.

В подразделе 1.1 «Классические и адаптивные методы стеганографии, метрики оценки качества» сформирован терминологический аппарат исследования. В рамках работы разграничены и детерминированы базовые понятия:

- Цифровая стеганография - как научно-техническое направление, сосредоточенное на сокрытии самого факта существования канала связи путем внедрения данных в цифровые носители;
- Стегоанализ - как дисциплина, направленная на обнаружение скрытого контента и оценку защищенности систем;
- Изображение-контейнер (Cover image) - как исходная среда для интеграции данных;
- Стеганограмма (Stego image) - как результирующий заполненный объект.

В данном подразделе проведена подробная классификация существующих подходов. Рассмотрены методы пространственной области, основанные на прямой модификации пикселей: алгоритм прямой замены наименьшего значащего бита (LSB-replacement) и более устойчивый алгоритм LSB-согласования (LSB-matching). Проанализированы методы частотной области, использующие спектральные преобразования (DCT, DWT), а также современные адаптивные алгоритмы (семейства UNIWARD, HUGO), динамически вычисляющие «стоимость искажений» для защиты наиболее уязвимых областей изображения.

На основе концепции «стеганографического треугольника» доказано, что традиционные методы не обеспечивают одновременной оптимизации емкости, скрытности и робастности трафика. Также в подразделе формализован

математический базис метрик оценки качества: пикового отношения сигнала к шуму (PSNR), индекса структурного сходства (SSIM), частоты битовых ошибок (BER), перцептивного критерия LPIPS и площади под ROC-кривой (AUC) для оценки статистической слепоты систем стегоанализа.

В подразделе 1.2 «Архитектура и принципы обучения нейросетевых стеганосистем на базе GAN» описан и теоретически обоснован переход к моделям глубокого обучения. Спроектирована и детально разобрана трехкомпонентная сквозная архитектура, функционирующая по созависимому принципу:

- Энкодер (Сеть-внедритель) - нейросетевой программный модуль, осуществляющий интеграцию сообщения в контейнер с минимизацией визуальных артефактов;
- Декодер (Сеть-извлекатель) - модуль, реализующий «слепое» (без наличия оригинала) побитовое восстановление скрытого сообщения из стеганограммы;
- Дискриминатор (Критик) - состязательный модуль, выполняющий роль автоматизированного стегоанализатора и оценивающий статистическую неотличимость распределений.

Особое внимание уделено математическому описанию комбинированной многокритериальной функции потерь общего вида:

Описана динамика обратного распространения ошибки (*backpropagation*), при которой градиенты от декодера и дискриминатора одновременно корректируют весовые коэффициенты энкодера, обеспечивая экспериментальный подбор баланса между точностью извлечения, визуальной невидимостью и устойчивостью к детекции.

В подразделе 1.3 «Обзор инструментальных средств и существующих решений» выполнен критический анализ программных фреймворков и библиотек общего назначения (PyTorch, TorchVision, Pillow), а также специализированных открытых нейросетевых стеганосистем, в частности

базовых моделей SteganoGAN и HiDDeN.

В ходе анализа были выявлены их ключевые деструктивные ограничения: высокая подверженность эффекту «коллапса моды» в процессе состязательной игры, отсутствие встроенных блоков пространственного внимания (что приводит к демаскирующему размытию гладких зон изображений), а также полное отсутствие встроенных механизмов долгосрочного сохранения и мониторинга промежуточных состояний экспериментов в реальном времени. На основе выявленных недостатков существующих решений сформирована и аргументирована необходимость проектирования собственной расширенной архитектуры стеганосистемы с гибкой системой конфигурирования.

Второй раздел «Практическая часть» посвящён проектированию, реализации и экспериментальной оценке разработанной системы.

В подразделе 2.1 задача формализована через функции встраивания и извлечения . Установлены три критерия эффективности: визуальная скрытность (), надёжность извлечения ( ) и статистическая неразличимость (точность детектора). Математическая формализация стеганографического процесса позволила перейти от эвристических правил к строгому описанию через функции встраивания и извлечения , где - изображение-контейнер размером , а - секретное двоичное сообщение длины Это создало основу для объективного сравнения разработанной системы с классическими методами стеганографии.

В подразделе 2.2 описана программная реализация на Python 3.10 / PyTorch в среде Google Colab. Разработанная сквозная архитектура включает четыре взаимодействующих компонента: DenseEncoder с механизмом пространственного внимания, DenseDecoder, критик WGAN-GP со спектральной нормализацией и стегоанализатор с банком SRM-фильтров.

Многокритериальная функция потерь генератора объединяет четыре компонента: MSE-потери качества изображения (), бинарную кросс-энтропию восстановления сообщения (), состязательные потери WGAN и инвертированные потери стегоанализатора. Такой баланс весов обеспечивает приоритет визуальной скрытности на ранних этапах обучения. Реализованная

система персистентности состояния эксперимента (run_state.json, history.json, metrics.jsonl) обеспечивает бесшовное продолжение обучения после принудительного разрыва сессии.

В подразделе 2.3 представлены результаты эксперимента run_03 (20 эпох, датасет MS COCO Val2017), в ходе которого были достигнуты следующие показатели: среднее значение дБ, , точность декодера = , а показатель accuracy состязательного критика стабилизировался в районе . Проведено ручное тестирование сквозного цикла кодирования–декодирования с верификацией устойчивости данных к сериализации в формат PNG.

Рисунок 1 - Сравнение оригинального изображения (Cover),
стеганограммы (Encoded) и карты разностей (Residual/Noise).

Анализ полученных карт разностей (рис. 2) наглядно подтвердил, что разработанный механизм пространственного внимания успешно распределяет скрытый сигнал исключительно по высокочастотным текстурированным областям и контурам объектов, полностью избегая внесения демаскирующих артефактов в гладкие и однородные зоны изображения.

Достигнутые показатели accuracy критика в диапазоне 0,52–0,54 свидетельствуют о том, что система вышла на уровень стабильного равновесия

GAN.

```
Inference device: cuda
Approx text capacity in bytes: 60
{
  "input_text": "Kotik",
  "decoded_from_tensor": "Kotik",
  "decoded_from_saved_png": "Kotik",
  "image_loss": 0.0010080165229737759,
  "saved_stego_path": "/content/drive/MyDrive/diploma_runs/run_01/manual_tests/manual_stego_fixed.png",
  "preview_path": "/content/drive/MyDrive/diploma_runs/run_01/manual_tests/manual_stego_fixed_preview.png",
  "repeat_factor": 64
}
```

Рисунок 2 – Результат извлечения встроенного сообщения в изображение

На рисунке 2 наглядно продемонстрирован интерфейс и результаты верификации разработанного программного комплекса в ходе сквозного ручного тестирования. Иллюстрация подтверждает корректность выполнения полного цикла обработки данных: от первоначального ввода текстового секретного сообщения и его скрытой интеграции в выбранное изображение-контейнер до сохранения готовой стеганограммы на диск.

ЗАКЛЮЧЕНИЕ

В ходе выполнения дипломной работы была полностью достигнута поставленная цель по разработке и исследованию стеганографической системы на базе генеративно-состязательной нейронной сети, обеспечивающей высокую степень скрытности и точность извлечения данных.

На первом этапе был проведён комплексный анализ классических и современных методов стеганографии, который выявил, что традиционные подходы вроде LSB или методов на основе частотных преобразований DCT и DWT остаются уязвимыми для статистического стеганоанализа, что подтвердило актуальность использования глубокого обучения. В рамках изучения архитектуры и принципов функционирования GAN было установлено, что состязательный характер обучения позволяет генератору адаптироваться к критериям дискриминатора, создавая визуально неотличимые от оригинала стеганограммы.

На основе полученных теоретических знаний была спроектирована и программно реализована с помощью библиотеки PyTorch трёхкомпонентная архитектура SteganoGAN, включающая сеть-внедритель, сеть-извлекатель и критик. Особое внимание при реализации уделялось балансировке весовых коэффициентов функций потерь – в частности, приоритету визуальной скрытности с весом 50,0 над точностью извлечения с весом 15,0.

Экспериментальное исследование эффективности системы в рамках заезда run_03 показало выдающиеся результаты по ключевым метрикам: среднее значение PSNR достигло 41,84 дБ, а индекс структурного сходства SSIM составил 0,996, что гарантирует отсутствие видимых артефактов даже при детальном анализе. При этом надёжность системы подтверждается точностью декодера на уровне 99,7%, что минимизирует вероятность ошибки при восстановлении скрытого сообщения.

Оценка устойчивости к стеганоанализу продемонстрировала, что точность критика в финале обучения стабилизировалась в районе 0,52. Это свидетельствует о достижении состояния равновесия, при котором

автоматизированные системы анализа не могут эффективно идентифицировать наличие скрытого контента.

Сравнительный анализ с классическими алгоритмами подтвердил превосходство предложенного нейросетевого подхода в части скрытности и устойчивости к обнаружению, что делает разработанную систему перспективным инструментом для защиты авторских прав и обеспечения конфиденциальности передачи информации в современных цифровых сетях. Задачи работы выполнены в полном объёме, что позволяет сделать вывод о высокой эффективности применения генеративно-состязательных сетей в области современной криптостеганографии.

ОСНОВНЫЕ ИСТОЧНИКИ ИНФОРМАЦИИ

1. Петрухин, А.В. Стеганография и методы защиты информации. – М.: Академия, 2019. – 224 с.
2. Al-Faydi, S.N.M., Ahmed, S.K. Improved LSB image steganography with high imperceptibility based on cover-stego matching // IET Image Processing. – 2023. – Vol. 17, No. 8. – P. 2031–2043. DOI: 10.1049/ipr2.12773.
3. Aghili, S., Jahangir, A. Hybrid machine learning combined with image segmentation for enhanced LSB steganography // Future Generation Computer Systems. – 2023. – Vol. 144. – P. 341–354. DOI: 10.1016/j.future.2023.03.013.
4. Holub, V., Fridrich, J. Designing steganographic distortion using directional filters // IEEE International Workshop on Information Forensics and Security (WIFS). – 2012. – P. 234–239.
5. Побыванец, В.В., Паюсова, Т.И. Разработка приложения для стеганографии на основе генеративно-состязательной нейронной сети // Вестник Тюменского государственного университета. – 2022. – № 4. – С. 305–309.
6. Барщевский, Е.Г. Генеративно-состязательные нейронные сети и их применение в изображениях // Информатика и компьютерная техника. – 2024. – № 2. – С. 45–53.