В. В. Побыванец, Т. И. Паюсова
Тюменский государственный университет, г. Тюмень
УДК 004.056

РАЗРАБОТКА ПРИЛОЖЕНИЯ ДЛЯ СТЕГАНОГРАФИИ НА ОСНОВЕ ГЕНЕРАТИВНО-СОСТЯЗАТЕЛЬНОЙ НЕЙРОННОЙ СЕТИ

Аннотация. *В статье рассмотрен процесс разработки приложения для стеганографии цифровых изображений на основе генеративно-состязательной нейронной сети.*

Ключевые слова: *стеганография изображений, генеративно-состязательные нейронные сети, искусственный интеллект, глубокое обучение, машинное обучение, нейросетевое моделирование.*

Введение. Проблема исследования. В настоящее время традиционная цифровая стеганография изображений (например, с применением методов LSB и Коха-Жао) сталкивается с проблемами обнаружения факта сокрытия данных современными стегоанализаторами. Метод стеганографии, построенный на использовании генеративно-состязательной нейронной сети (Generative Adversarial Networks, GAN), позволит скрывать сообщения в изображения таким образом, что стегоанализаторы не смогут найти в этих изображениях никаких намеков на манипуляции.

Помимо этого, область обучения генеративно-состязательных сетей активно развивается, что подтверждается количеством работ, выпускаемых ежегодно (см. рис. 1) [1].

Материалы и методы

Архитектура модели генеративно-состязательной нейронной сети

Основой для архитектуры нейронной сети стала генеративно-состязательная сеть Вассерштейна [2], состоящая из кодировщика, декодировщика и критика. Структура модели нейронной сети, используемой в приложении, изображена на рис. 2.

Кодировщик получает на вход изображение и секретное сообщение.

Процесс работы кодировщика выглядит следующим образом:

1. Полученное изображение раскладывается на тензор с 3 слоями.
2. Тензор секретного сообщения и изображения объединяются.

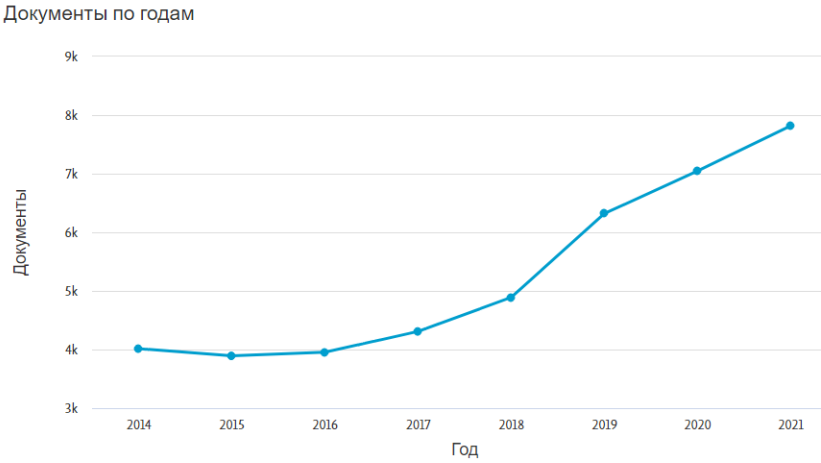

Рис. 1. Количество научных работ в области обучения генеративно-состязательных сетей (по данным Scopus)

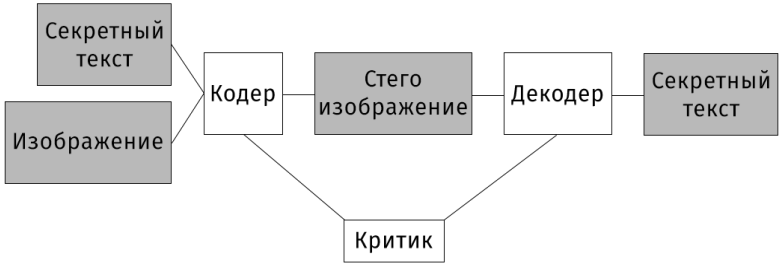

Рис. 2. Структура модели нейронной сети

3. Из полученного тензора генерируется изображение.

4. Происходит конкатенация сгенерированного изображения и оригинального изображения. Данное действие позволяет получить стегоизображение, максимально похожее на оригинал.

Результат работы кодировщика — стеганографическое изображение со встроенным в него секретным сообщением.

Декодировщик расшифровывает стеганографическое изображение. Он пытается восстановить секретный текст, подаваемый ранее на вход кодировщика, как можно точнее.

Критик позволяет оценить работу всей нейронной сети, путем дачи оценки работе кодировщика и декодировщика.

Обучение

Обучение производилось на тренировочном и валидационном наборах данных, состоящих из случайно выбранных изображений и сгенерированных битов данных для сокрытия.

Изображения для наборов были объединены из двух датасетов — COCO и Div2k. Данные датасеты предоставляют большой выбор изображений различного разрешения и тематики, что в перспективе предотвратит проблему исчезающего градиента при обучении.

Модель кодер-декодер для обучения получала на вход тренировочный набор данных. После тренировки на вход модели поступал валидационный набор данных для последующего расчета функций потерь.

Для отслеживания качества обучения и оптимизации модели были использованы следующие функции потерь: среднеквадратичная ошибка, кросс-энтропия и функция потерь Вассерштейна.

Для оптимизации декодировщика, использовалась кросс-энтропийная функция потерь, выражающаяся формулой (1).

$$I(p,q) = -\sum p(x)\, log\, q(x), \tag{1}$$

где x — прогнозируемые результаты работы нейронной сети, $p(x)$ — распределение вероятной истинной выборки, $q(x)$ — распределение вероятностей прогноза нейронной сети.

Функция измеряла разницу между исходным секретным сообщением и сообщением, извлеченным декодировщиком из стегоизображения.

Для оптимизации кодировщика использовалась среднеквадратичная ошибка, выражающаяся формулой (2).

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y})^2, \tag{2}$$

где N — размер выборки, y_i — прогноз вероятности от нейронной сети, $\hat{y}$ — истинное значение.

Функция сравнивала каждый пиксель сгенерированного стегоизображения и оригинального изображения для оценки их схожести.

Для оптимизации работы критика использовалась функция потерь Вассерштейна, выражающаяся формулой (3).

$$W(y,p) = -\frac{1}{n}\sum_{i=1}^{n}(y_i p_i), \quad (3)$$

где n — размер выборки, y_i — прогноз вероятности от нейронной сети, p_i — истинное значение.

Цель обучения — минимизировать сумму этих трех потерь.

Результаты. Результат работы приложения для стеганографии на основе обученной модели генеративно-состязательной нейронной сети изображен рис. 3.

Рис. 3. Результат работы нейронной сети

Все изображения, полученные в процессе обучения нейронной сети и подготовки данного материала, были проверены в стегоанализаторах.

В качестве программ для анализа были выбраны StegoAnalyzer и StegExpose [3].

По результатам проверки, никакое стегоизображение не было помечено в качестве подозрительного. Это означает, что метод стеганографии, основанный на нейронных сетях, не может быть обнаружен простыми стегоанализаторами.

Заключение. Модель генеративно-состязательной нейронной сети, рассмотренная в данной работе, способна сокрыть до 4 бит информации на пиксель изображения любого разрешения. Сокрытие в изображение данных свыше 4 бит на пиксель может привести к ошибке или неправильному декодированию секретного сообщения.

Визуальное отличие обработанного кодировщиком изображения от оригинала заключается в небольшом, малозаметном шуме на открытых участках изображения.

Полученное стегоизображение не является подозрительным для современных стегоанализаторов, а извлечение секретного сообщения не является возможным.

В дальнейшем, приложение будет дополнено модулем криптографии, поскольку это усилит защищенность секретного сообщения без вреда для качества стегоизображения. Секретное сообщение, перед попаданием в кодировщик, будет зашифровано при помощи закрытого ключа.

СПИСОК ЛИТЕРАТУРЫ

1. Библиографическая и реферативная база данных и инструмент для отслеживания цитируемости статей, опубликованных в научных изданиях. — Текст : электронный // Scopus : [сайт]. — URL: https://www. scopus.com/ (дата обращения: 24.05.2022).
2. Wasserstein GAN. Arxiv : [сайт]. — URL: https://arxiv.org/abs/1701.07875 (дата обращения: 24.05.2022). — Text : electronic.
3. StegExpose — A Tool for Detecting LSB Steganography : [сайт]. — URL: https://www.embeddedsw.net (дата обращения: 24.05.2022). — Text : electronic.