# ИНФОРМАЦИОННАЯ БЕЗОПАСНОСТЬ

УДК 519.85 ISSN 1995-5499
DOI: https://doi.org/10.17308/sait.2021.1/3369
Поступила в редакцию 19.03.2021
Подписана в печать 26.04.2021

# СТЕГОАНАЛИЗ ЦИФРОВЫХ ИЗОБРАЖЕНИЙ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ ПОВЕРХНОСТНОГО И ГЛУБОКОГО МАШИННОГО ОБУЧЕНИЯ: ИЗВЕСТНЫЕ ПОДХОДЫ И НОВЫЕ РЕШЕНИЯ

**© 2021 А. А. Сирота, М. А. Дрюченко, А. Ю. Иванков**✉

*Воронежский государственный университет*
*Университетская пл., 1, 394018 Воронеж, Российская Федерация*

**Аннотация.** Рассматривается современное состояние проблемы стегоанализа цифровых изображений, направленной на исследование и разработку эффективных методов выявления стеганографически скрытых (визуально незаметных) сообщений в контейнерах-изображениях. В первой части статьи проводится общая классификация известных подходов и детальный обзор ранее полученных результатов в области стегоанализа на основе использования методологии поверхностного (shallow learning) и глубокого машинного обучения (deep learning). Описываются используемые в современных системах поверхностного машинного обучения системы признаков и реализуемые на их основе классификаторы (композиционные алгоритмы, алгоритмы на основе метода опорных векторов и др.). В качестве альтернативы рассматриваются возможности глубоких нейронных сетей с архитектурами, реализующими, в основном, сверточную обработку с различными модификациями (дополнительные слои обработки, функции активации специального вида и т.п.). Приводятся данные сравнительного анализа эффективности применения альтернативных подходов и архитектур нейронных сетей, применяемых для решения задач стегоанализа. Сравнение проводится для стандартных наборов изображений применительно к использованию при внедрении стегосообщений методов адаптивной пространственной стеганографии WOW, HUGO, S-UNIWARD. Отмечается высокая степень универсальности и эффективности глубокого машинного обучения как перспективного направления развития методологии стегоанализа. Во второй части статьи описывается предложенная авторами архитектура глубокой нейронной сети и результаты ее применения в задачах стегоанализа цветных изображений. Общей идеей реализуемого в этой части работы подхода является использование относительно простых сверточных сетей для последовательного анализа небольших фрагментов (блоков) исходных больших изображений с объединением получаемых результатов классификации как последовательности бинарных признаков по схеме наивного байесовского классификатора. Исследования проведены для базы цветных изображений PPG-LIRMM-COLOR database и алгоритмов WOW и S-UNIWARD, используемых для внедрения стегосообщений при различных объемах полезной нагрузки. Показано, что получаемая точность стегоанализа изображений большого размера сопоставима с результатами, полученными ранее другими авторами, а, в некоторых случаях, и превосходит их.

**Ключевые слова:** стеганография, стегоанализ, стегосообщение, цифровые изображения, машинное обучение, глубокие нейронные сети.

✉ Иванков А. Ю.
e-mail: ivankov@cs.vsu.ru

## ВВЕДЕНИЕ

Задача стегоанализа (СА), как известно [1–3], состоит в обнаружении (выявлении) факта внедрения визуально незаметного стегосообщения (цифрового водяного знака) в анализируемый объект цифрового контента (изображение, видео, звуковой сигнал и т. п.) и оценке параметров внедренного сообщения. Обычно в качестве модели стеганографически скрытой информации (ССИ) рассматривается псевдослучайная двоичная (битовая) последовательность. Задача СА может решаться в прямой постановке, как задача анализа контейнера с неизвестным содержимым, так и как обратная задача проверки скрытности разрабатываемых алгоритмов компьютерной стеганографии по отношению к стегоанализу (анализ за «противника»).

Об интенсивности исследований и разработок в этой области свидетельствует наличие большого количества программных средств стегоанализа. В настоящий момент существует более 100 различных стеганографических программных пакетов, большая часть которых ориентирована на анализ изображений. В последние 10–15 лет развитие стегоанализа осуществляется на основе использования методов и алгоритмов машинного обучения как универсального и потенциально эффективного подхода к решению любых задач анализа данных различной природы. При этом разнообразие применяемых решений очень велико, что требует проведения систематизированного анализа имеющихся источников с целью определения перспективных направлений исследований и разработок.

Поэтому целью данной работы является проведение обзора известных публикаций, как иностранных, так и отечественных по этой теме, а также изложение оригинальных результатов авторов в плане использования глубоких нейронных сетей сверточного типа для обработки цветных изображений большого размера и их сравнение с результатами, полученными другими авторами.

## 1. ОБЗОР ИЗВЕСТНЫХ РАБОТ В ОБЛАСТИ СТЕГОАНАЛИЗА

В литературе обычно используют следующую классификацию методов и алгоритмов СА.

1. По степени универсальности используемых методов и анализируемых контейнеров выделяют:

специализированные (не «слепые») методы и алгоритмы, предназначенные для выявления сообщений, внедренных определенным алгоритмом ССИ;

универсальные («слепые») методы и алгоритмы, предназначенные для выявления сообщений без привязки к определенным алгоритмам ССИ.

2. По влиянию на контейнер, потенциально содержащий цифровой водяной знак (ЦВЗ), методы классифицируются таким образом:

пассивные методы, которые направлены на определение только факта наличия или отсутствия ранее внедренных сообщений;

активные методы стегоанализа, которые определяют параметры внедренного сообщения и используемого алгоритма и обеспечивают, по возможности, его расшифровку или устранение (уничтожение).

3. По используемым принципам обработки информации в ходе стегоанализа большинство известных методов разделяется следующим образом:

сигнатурные методы, в основе которых лежит визуальный (качественный) или синтаксический количественный анализ последовательности символов, представляющих контейнер, на предмет наличия не характерных включений;

вероятностные методы анализа контейнера, которые направлены на обнаружение отклонений используемых статистик (гистограммы, непараметрические статистики различных видов), от статистик, характерных для «естественных» (не подвергнутых модификации при внедрении сообщения) объектов;

методы машинного обучения, которые базируются на построении классификаторов объектов для выявления внедренных сообщений, с использованием представительных

обучающих выборок, содержащих заполненные и незаполненные контейнеры.

Последний класс методов в настоящее время развивается достаточно интенсивно и, в ряде случаев, показывает весьма высокую эффективность по отношению к наиболее скрытным алгоритмам ССИ. В большинстве ситуаций суть этого подхода, как уже упоминалось, состоит в построении классификатора объектов-контейнеров для обнаружения факта наличия ССИ на основе реализации процедуры обучения по представительным наборам обучающих примеров (обучающих выборок). Отличительной особенностью подобного подхода является, прежде всего, его универсальность. В дальнейшем мы остановимся на представлении авторских результатов именно в этой области. Кроме того, далее мы ограничимся рассмотрением исключительно задачи выявления факта внедрения ЦВЗ, причем, без существенной потери общности используемых методов, по отношению к цифровым контейнерам-изображениям.

Методы и реализующие их алгоритмы машинного обучения в приложении к задаче стегоанализа можно разделить на две большие группы:

классические «неглубокие» методы и алгоритмы машинного обучения или методы поверхностного обучения (shallow methods);

методы и алгоритмы, основанные на применении глубоких нейронных сетей (deep learning methods).

В соответствии с этим разделением и будет проводиться далее рассмотрение известных результатов в области применения методов машинного обучения для стегоанализа изображений и других объектов цифрового контента.

### 1.1. Применение неглубоких методов машинного обучения

К классическим неглубоким методам и реализующим их алгоритмам машинного обучения относятся такие как: наивный байесовский классификатор; метод опорных векторов; композиционные алгоритмы на основе бэггинга («случайный лес» и его модификации); композиционные алгоритмы на основе бустинга (Adaboost и его модификации) и ряд других. Характерной особенностью этих алгоритмов (в отличие от алгоритмов глубокого обучения) является необходимость предварительной обработки анализируемых объектов для извлечения совокупности информативных признаков, используемых при обучении классификаторов. Обширный обзор публикаций, иллюстрирующий широту перечня применяемых методов и алгоритмов стегоанализа, включая и указанные методы машинного обучения, приведен в работе [4].

Одними из первых в 2002 году Сьюви Лью и Хани Фарид представили новое направление в стегоанализе — метод машинного обучения [5–7]. Появление этого направления можно рассматривать как ответную меру на появление новых эффективных и скрытных алгоритмов ССИ, основанных на минимизации вносимого внедрением искажения, таких, например, как Outguess и F5. Предложенный ими подход заключался в использовании широко применяемых в машинном обучении алгоритмов, построенных на основе метода опорных векторов. В качестве исходного набора признаков в таких алгоритмах используется вектор, вычисляемый из статистических характеристик распределения групп пикселей изображения: математическое ожидание, дисперсия, среднеквадратичное отклонение и т. д. В этих работах обучение проводилось по выборке из 1800 пустых контейнеров и случайного подмножества из 1800 заполненных контейнеров. Использовался 72-х размерный вектор признаков, в результате чего получены следующие результаты: для алгоритма LSB точность классификации составила 99 %, для алгоритма Outguess около 95 %.

В развитие этого подхода для решения задачи стегоанализа в последующем разработаны специальные многомерные (порядка $10^3 \ldots 10^4$) системы (пространства) признаков. К наиболее часто используемым относятся следующие:

система SPAM (Subtractive Pixel Adjacency Matrix) [8], базирующая на вычислении ма-

трицы разностей смежных значений пикселей с последующим выделением набора производных признаков;

система SRM (Spatial Rich Model) [9], являющаяся одним из наиболее распространенных набором признаков для проведения стегоанализа, в которой чтобы отразить изменения в корреляции пикселей, используются линейные и нелинейные фильтры с различными ядрами и анализируются их отклики;

система PSRM (Projection Spatial Rich Model) [10], являющаяся усовершенствованным вариантом SRM, позволяющим несколько улучшить результаты, но более сложная и затратная в вычислительном отношении.

Указанные признаковые системы имеют многочисленные модификации и постоянно совершенствуются.

Эффективность различных алгоритмов обработки информации для обнаружения скрытых сообщений обычно демонстрируется на черно-белых изображениях размером 512×512 из стандартизованной базы BOSSBase 1.01 [11], часто используемой специалистами по стеганографии и стегоанализу. Для цветных изображений часто используется стандартизированная база 512×512 PPG-LIRMM-COLOR database [12].

Для проверки и сравнения алгоритмов стегоанализа и систем признаков проводится создание обучающих и тестовых примеров на основе наиболее скрытных алгоритмов встраивания ЦВЗ. Для контейнеров-изображений в этой постановке подлежащая обнаружению информация обычно внедряется при помощи современных методов адаптивной стеганографии HUGO, S-UNIWARD и WOW [13-15]. Эти методы считаются наиболее трудно обнаружимыми на данный момент и именно по ним приведены лучшие из известных результатов обнаружения факта наличия или отсутствия ССИ. [10]. Идея адаптивного внедрения заключается в том, что позиции для внедрения выбираются не произвольно, а исходя из свойств изображения; при этом с большей вероятностью внедрение осуществляется в те области, где обнаружить информацию должно быть труднее. Реже рассматриваются и другие классические и современные алгоритмы ССИ, реализуемые как в пространственной, так и в частотной области.

В ходе последних исследований и сравнительного анализа различных алгоритмов, показано, что наибольшей эффективностью, как и в более ранних исследованиях, обладают алгоритмы, основанные на использовании метода опорных векторов. Несколько хуже, но также часто достаточно эффективно работают ансамблевые (композиционные) алгоритмы на основе бустинга и бэггинга (AdaBoost и Random Forest).

Одним из эффективных приемов для стегоанализа цифровых изображений является использование алгоритмов сжатия в различных постановках. В этом плане в работах отечественных авторов [16, 17] предложен метод, который может использоваться как для не «слепых», так и для «слепых» алгоритмов обнаружения ССИ, идея которого основана на том, что включаемые данные статистически независимы от контейнера. При добавлении скрытых данных в контейнер его размер при сжатии определенным образом возрастает по сравнению с размером при сжатии исходного «пустого» контейнера, что может быть эффективно использовано при стегоанализе на основе простого порогового алгоритма.

В работе [18] используется понятие интегрального классификатора, состоящего из набора отдельных классификаторов, каждый из которых обрабатывает только те контейнеры, которые предварительно автоматически отфильтрованы для него. Теоретически интегральный классификатор может быть реализован разными способами. В указанной работе предложен интегральный классификатор на основе сжатия данных. Обучающее множество разбивается на несколько частей в соответствии с коэффициентом их сжатия и после этого обучается соответствующее количество классификаторов, причем каждый из них обучается на своем подмножестве. Во время тестирования контрольного множества очередной контейнер отправляется на классификатор, обученный на контейнерах, коэффициент сжатия которых наиболее близок к коэффициенту сжатия данного контейнера. Идея использования коэффициента

сжатия как критерия выбора классификатора возникла на основе известного факта, состоящего в том, что проводить стегоанализ шумных изображений сложнее, чем изображений с большими областями приблизительно одного цвета. Как известно, изображения первого типа сжимаются хуже, чем второго. Соответственно, классификаторы для плохо сжимаемых контейнеров должны обучаться на плохо сжимаемых контейнерах. Аналогично и для хорошо сжимаемых контейнеров. Принципиальное отличие этого подхода от других подходов, использующих сжатие данных, заключается в том, что здесь сжатие используется на предварительном этапе выбора классификатора, но не для построения самого алгоритма стегоанализа.

Обобщенное представление результатов исследований по применению алгоритмов машинного обучения для СА полутоновых изображений с использованием указанных выше наборов признаков можно получить, например, из работы [18]. Здесь дано сравнение результатов с использованием предложенного авторами интегрального классификатора со сжатием данных с результатами работы ансамбля классификаторов в виде простых линейных дискриминаторов, обучаемых по случайно формируемым подвыборкам, приведенных в более ранних работах [10, 19]. Данные по этим сравнительным результатам даны в цитируемой из [18] сводной табл. 1.

В дополнение к этим данным следует заметить, что использование набора признаков PSRM в экспериментах [10] лишь незначительно снижает ошибки на 1…2 %. Представленные в таблице цифры могут служить ориентирами для оценки достижимой точности обнаружения ССИ в черно-белых изображениях с использованием стандартных алгоритмов машинного обучения.

В области анализа цветных изображений также проведены соответствующие исследования. Для обучения классификаторов здесь также изначально используются системы признаков под общим названием Color Rich Model [20], которые базируются на адаптации SPAM к цветным изображениям. В работе [21] Color Rich Model дополнена признаками, полученными на основе анализа корреляции градиентов цветовых компонент R,G,B. В работе [22] использована так называемая гибридная система признаков, объединяющая основные характеристики модели, использующей матрицу совместной встречаемости уровней серого (GLCM, Gray-Level

*Таблица 1. Результаты сравнения точности обнаружения различных алгоритмов ССИ на множестве изображений BOSSBase 1.01 при различных объемах внедрения pl в %*
*[Table 1. Results of comparing of various algorithms by the detection accuracy of steganographically hidden information on a set of BOSSBase 1.01 images at various embedding volumes pl in %]*

| Анализируемый алгоритм ССИ | WOW с различным параметром загрузки контейнера | | | HUGO с различным параметром загрузки контейнера | | | S-UNIWARD с различным параметром загрузки контейнера | | |
|---|---|---|---|---|---|---|---|---|---|
| Вид классификатора | $pl = 0.1$ | $pl = 0.2$ | $pl = 0.4$ | $pl = 0.1$ | $pl = 0.2$ | $pl = 0.4$ | $pl = 0.1$ | $pl = 0.2$ | $pl = 0.4$ |
| Ансамблевый классификатор + SRM [10] | 62 | 69 | 81 | 65 | 77 | 89 | 59 | 69 | 80 |
| Единичный классификатор SVM+SRM [19] | 62 | 71 | 79 | 65 | 73 | 85 | 63 | 70 | 83 |
| Интегральный классификатор [19] | 76 | 87 | 92 | 76 | 87 | 92 | 75 | 85 | 94 |

Co-Occurrence Matrix), рассчитанных для отдельных цветовых каналов и их битовых плоскостей. Кроме того, предлагается использовать расширенный набор функций от GLCM в котором рассматриваются не только соседние пиксели, но и пиксели, расположенные на различных расстояниях от опорного.

Показательными являются результаты работы [21]. Здесь мы видим результаты для алгоритмов WOW и S-UNIWARD, причем в ситуации, когда встраивание проводится только в один (зеленый) канал и равномерно во все цветовые каналы. Следует отметить, что полученные результаты незначительно улучшают результаты на основе Color Rich Model, полученные в [20]. Как утверждают авторы, при загрузке только одного цветового канала точность выявления ССИ в контейнере выше, чем при одновременной загрузке всех каналов. При этом данные по одному каналу близки с лучшими результатами, полученными для черно-белых изображений.

В целом следует отметить, что результаты, полученные в современных исследованиях по стегоанализу на основе разнообразных систем многомерных признаков, демонстрируют достаточно высокие показатели точности классификации пустых и заполненных контейнеров.

### 1.2. Применение глубоких нейронных сетей

Как и в случае использования алгоритмов shallow learning, при использовании методов глубокого обучения задача состоит в обучении бинарного классификатора (классификатора на два класса) для выявления факта скрытия данных (ЦВЗ, сообщения) в анализируемом контейнере. При этом в подавляющем большинстве работ рассматривается задача стегоанализа контейнеров-изображений и сверточные нейронные сети (CNN) в различных модификациях и усовершенствованиях. Отличный англоязычный обзор на эту тему можно увидеть в работе [23].

Одной из первых работ в этом направлении является работа [24]. В этой работе авторы предложили специализированную архитектуру сверточной нейронной сети, которую они назвали CNN model called Gaussian-Neuron (GNCNN). Ее особенностью являлось использование пространственного фильтра высоких частот с фиксированным ядром, специальных функций активации в виде гауссианы, центрированной относительно нулевого значения входа, и слоев субдискретизации с усреднением в пределах окна пуллинга (Average Pooling) вместо часто используемого слоя (Max Pooling). На вход сети подавались черно-белые изображения размером 256×256, подвергнутые обработке высокочастотным фильтром размером 5×5. Первый сверточный слой фильтрует вход с ядром размером 5×5. Второй сверточный слой принимает выходные данные первого слоя в качестве входных данных и фильтрует его 16 ядрами размера 5×5. Третий, четвертый и пятый сверточные слои применяют свертки с 16 ядрами размера 3×3 соответственно, а шестой сверточный слой — с 16 ядрами размера 5×5. Активация в виде гауссианы применяется к каждому выходному сигналу, начиная со второго по шестой сверточные слои. Каждый сверточный слой сопровождается пуллингом размера 3×3 и с шагом 2, который работает с картой признаков в соответствующем сверточном слое, что приводит к итоговому понижению размерности извлекаемых признаков до 256.

Извлеченные признаки передаются модулю классификации, который состоит из трех полносвязных слоев. Выход каждого нейрона в первых двух полносвязных слоях GNCNN активируются обычной функцией Relu. Последний полносвязный слой имеет два нейрона, и его выходной сигнал подается на вход бинарного классификатора, реализуемого с использованием активации Softmax.

Главная идея такой обработки состоит в том, что пространственный фильтр локализует малые искажения в областях исходного контейнера, связанные с внедрением ССИ. Высокочастотный стегошум, добавленный к исходному изображению, представляет собой очень слабый сигнал, на который сильно влияет содержание изображения. Следовательно, с помощью фильтрации верхних частот можно усилить слабый стегосигнал и уменьшить влияние исходного контента. Использование

гауссовской активационной функции обеспечивает преимущественную реакцию сверточных слоев сети на этот стегосигнал, значения которого локализованы в окрестности нуля, и подавление входных воздействий, вызванных прохождением через фильтр отдельных участков изображения.

В итоге, авторам удалось получить следующие результаты по отношению к уже упоминавшимся алгоритмам адаптивной стеганографии HUGO, S-UNIWARD и WOW. Для BOSSBase во всех трех случаях внедрения ССИ с различной полезной нагрузкой GNCNN обеспечивает гораздо меньшую ошибку обнаружения, чем при использовании SVM с гауссовским ядром и признаковой системой SPAM. По сравнению с использованием набора SRM ансамбля классификаторов, ошибка оказалось примерно на 2…5 % выше в зависимости от уровня полезной нагрузки. Эксперименты на базе ImageNet показывают, что GNCNN достигает ошибки обнаружения, близкой по отношению к классификатору с набором признаков SRM. В последующей работе этих авторов [25] рассматриваются возможности извлечения и анализа карт признаков для стегоанализа, формируемых при обучении глубоких сетей класса CNN.

В следующей по времени работе [26] рассматривается новая архитектура сверточной нейронной сети, содержащая всего два сверточных слоя, причем первый слой здесь выполняет роль настраиваемого фильтра предобработки, а второй имеет 64 канала и большую размерность ядра, сопоставимую с размерностью обрабатываемого изображения, выполняя при этом фактически функцию нескольких полносвязных слоев. В слоях свертки используются функции активации в виде гиперболического тангенса (Tanh). Завершает обработку слой классификации на два класса с активацией Softmax. Полученные результаты (аналогично на черно-белых изображениях) показывают существенное повышения качества обнаружения ССИ, при котором значения точности классификации достигает для HUGO, S-UNIWARD и WOW величин порядка 80…97 % в зависимости от объема полезной нагрузки. Однако выводы, сделанные в этой работе, ограничены использованием одного и того же ключа встраивания стегосообщения в разные изображения (см. также работу [27]). В работе Н. А. Нагорного [28] проведена проверка работы такой сети и также показаны преимущества данной архитектуры по сравнению с GNCNN. При этом предложен комбинированный вариант двуслойной сети, в который дополнительно введен высокочастотный пространственный фильтр предобработки, используемый в GNCNN.

В последних по времени публикациях, посвященных использованию методологии Deep Learning, в СА используются самые разнообразные по архитектуре глубокие сети [23]. Из этих сетей как наиболее эффективную следует выделить сеть Yedrouj-Net [29]. Ее архитектура предполагает использование шести сверточных и трех полносвязных слоев. Кроме того, в первых сверточных слоях используются нелинейные активации в виде функций Abs(…) и Trunk(…) (линейные функции с ограничением по порогу снизу и сверху). Полученные с использованием такой сети точности классификации при анализе данных из базы BOSSBase имеют значения порядка 72…86 % и превосходят результаты, демонстрируемые другими архитектурами (сети Xu-Net, Ye-Net [23]).

Сравнительные результаты для представленной архитектуры Yedrouj-Net по отношению к сетям Xu-Net, Ye-Net, а также ансамблю классификаторов в сочетании с набором признаков SRM (EC+SRM) представлены в табл. 2.

В работе отечественных авторов 2020 года [30] проводится исследование собственной модели сверточной сети и анализ известных результатов по использованию подобных алгоритмов в сравнении с алгоритмами неглубокого машинного обучения. В ходе выполненных экспериментов авторам на основе анализа сравнительно небольшого количества изображений удалось получить точность классификации порядка 85%, что сопоставимо с ранее полученными результатами. Одновременно проводится количественный анализ различных неглубоких классификаторов и систем признаков и нейросетевых решений.

*Таблица 2. Результаты сравнения ошибок обнаружения для глубоких нейронных сетей при различных объемах внедрения (полезной нагрузке) в %*
*[Table 2. Results of comparing detection errors for deep neural networks with various embedding volumes (payload) in %]*

| Анализируемый алгоритм ССИ | WOW с различным параметром загрузки контейнера | | S-UNIWARD с различным параметром загрузки контейнера | |
|---|---|---|---|---|
| | $pl = 0.2$ | $pl = 0.4$ | $pl = 0.2$ | $pl = 0.4$ |
| EC+SRM | 63 | 74 | 63 | 75 |
| Yedrouj-Net | 72 | 86 | 63 | 77 |
| Xu-Net | 67 | 79 | 61 | 73 |
| Ye-Net | 67 | 77 | 60 | 69 |

Отмечается, что существенным недостатком статистических классификаторов, отсутствующим в методах на основе нейронных сетей, является их узкая специализация на строго определенных методах формирования стего-контейнеров.

Таким образом, можно констатировать, что в настоящее время наиболее значительный объем исследований и разработок в области стегоанализа обеспечивают исследования в области обучения глубоких нейронных сетей для построения классификаторов цифровых контейнеров (прежде всего изображений) как универсального и потенциально весьма эффективного подхода.

## 2. МЕТОДЫ И МАТЕРИАЛЫ

В ходе своих исследований авторы поставили задачу разработки алгоритмов стегоанализа на основе глубоких сверточных сетей, ориентированных на анализ цветных изображений большого размера (512×512 и более). Идея предлагаемого подхода состоит в проведении обучения сетей относительно простой архитектуры на небольших фрагментах (блоках) размером 32×32, 64×64, 128×128 для реализации последовательной вторичной обработки совокупности результатов классификации, выполненной с использованием ранее обученного нейросетевого классификатора на блоках целостного изображения большого размера, с целью принятия окончательного решения.

Полученные результаты бинарной классификации обрабатываются по схеме наивного байесовского классификатора, что фактически сводится к сравнению количества «положительных» и «отрицательных» ответов с пороговым значением, зависящим от общего количества блоков в целостном изображении.

Можно выдвинуть два аргумента в пользу такой обработки. Во-первых, применение алгоритмов адаптивной стеганографии при ССИ, таких, например, как WOW, HUGO, S-UNIWARD, предполагает неравномерное встраивание стегосообщения в пространственные области контейнера. Используются, в основном, зашумленные области, для которых вносимые искажения минимальны. Во-вторых, можно ожидать, что обучение нейронных сетей для анализа небольших фрагментов окажется не таким затратным и можно реализовать более простые архитектуры. Собственно, проверке этих положений и посвящена вторая часть данной работы.

### 2.1. Предлагаемая архитектура нейронной сети

В ходе многочисленных экспериментов нами была предложена архитектура сверточной нейронной сети, отличающаяся от известных и содержащая всего три обучаемых сверточных слоя и три обучаемых полносвязных слоя.

В сети дополнительно введен слой предобработки входного изображения (для всех трех каналов), в котором реализован пространственный высокочастотный фильтр с возможностью гибкой перестройки величины и параметров фильтра. Ядро фильтра яв-

ляется симметричным и описывается следующим выражением:

$$h(x,y)=\begin{cases} I(x,y)E(x,y)S(x,y), & |x|\le\frac{d}{2},|y|\le\frac{d}{2},\\ 0, & |x|>\frac{d}{2},|y|>\frac{d}{2}, \end{cases}$$

$$I(x,y)=(-1)^{|x|+|y|+1},$$

$$E(x,y)=\exp\left[-\frac{1}{2}\alpha\sqrt{x^2+y^2}\right],$$

$$S(x,y)=\operatorname{sinc}(x/3)\operatorname{sinc}(y/3),$$

где $x, y$ — целочисленные значения аргумента, принимающие положительные и отрицательные значения в пределах размера ядра $d$. На рис. 1 представлено изображение функции ядра для $d=7,\ \alpha=0.25$. В наших экспериментах далее использовались параметры $d=5,\ \alpha=0.1$.

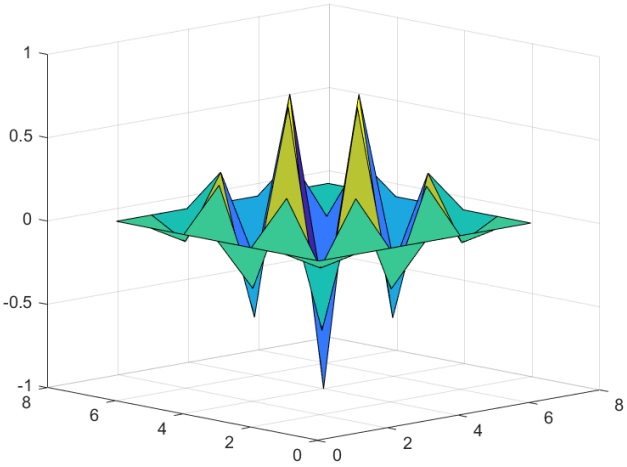


*Рис. 1. Типовой вид ядра высокочастотного пространственного фильтра*
*[Fig. 1. Typical view of the high-pass spatial filter kernel]*

Что касается используемых в сверточных слоях функций активации, то в первом сверточном слое после слоя высокочастотной фильтрации, свертки и батч-нормализации реализована (по аналогии с GNCNN) гауссовская функция активации с настраиваемым в процессе обучения параметром влияния $\sigma$ (среднеквадратичным отклонением). Начальная инициализация значения $\sigma$ проводится датчиком случайных чисел по равновероятному закону в диапазоне 0.01…0.5. Во всех остальных сверточных и полносвязных слоях используются активации Relu, за исключением последнего полносвязного слоя, на выходе которого используется стандартная активация Softmax. Следует отметить, что допустимо использование в первом слое и функции Relu, однако получаемые результаты оказываются несколько хуже.

Итоговая архитектура предложенной сети представлена на рис. 2.


*Рис. 2. Предлагаемая архитектура сверточной нейронной сети*
*[Fig.2. Proposed convolutional neural network architecture]*

На рис. 2 использованы следующие обозначения:

Image — входной слой $n_x \times m_y \times 3$, принимающий блок изображения в трех каналах цветности;

Pred_Hf — слой, отвечающий за выполнение пространственной высокочастотной фильтрации;

Conv_1 — первый сверточный слой с ядром свертки 5×5, шагом 1, и 32 каналами;

Conv_2, Conv_3 — второй и третий слои свертки с ядром 3×3, шагом 1, имеющие 64 и 128 каналов соответственно;

Bn_1, Bn_2, Bn_3 — слои стандартной батч-нормализации;

Gaussian_1 — гауссовская функция активации для первого слоя свертки $f(x)=\exp(-x^2/2\sigma^2)$;

Relu_2, Relu_3 — стандартные функции активации для второго и третьего слоев свертки типа «линейная с ограничением снизу» $f(x)=\max(0,x)$;

Av-1, Av-2, Av-3 — слои субдискретизации (пуллинга) на основе вычисления среднего (average pooling) в окне 3×3 с шагом 2;

Fc_1, Fc_2, Fc_3 — полносвязные слои, имеющие, соответственно 128, 128 и 2 выхода;

Relu_Fc_1, Relu_Fc_2 — стандартные функции активации для первого и второго полносвязных слоев свертки типа «линейная с ограничением снизу» $f(x)=\max(0,x)$;

Softmax — стандартная функция активации для классификации на два класса с использованием в качестве функции потерь кросс-энтропии;

Class — слой классификации, отвечающий за вычисление функции потерь кросс-энтропии при классификации на несколько взаимоисключающих классов образов.

Следует отметить, что увеличение количества сверточных слоев, а также использование слоев dropout, как показали многочисленные эксперименты, не дает ощутимого прироста точности классификации.

### 2.2. Алгоритм вторичной обработки результатов классификации

При анализе цветных изображений пространственного размера $N_x \times M_y$ в ходе вторичной обработки результатов классификации глубокой сетью совокупности блоков пространственного размера $n_x \times m_y$, размещаемых без перекрытия, получим в общей сложности $P = (N_x \times M_y)/(n_x \times m_y)$ бинарных ответов (считаем, что блоки по осям укладываются в исходном изображении кратное число раз).

В статистическом смысле можно считать получаемые по каждому блоку решения независимыми. Обозначим эти решения как $x_k$, $x_k \in \{1;0\}$, $k = \overline{1,P}$, где значение $x_k = 1$ обозначает решение в пользу потенциального наличия стегосообщения в анализируемом фрагменте изображения, а значение $x_k = 0$ обозначает решение в пользу отсутствия в нем стегосообщения. Для синтеза окончательного решающего правила по результатам вторичной обработки введем следующие обозначения для вероятностей значений бинарных признаков двух классов $\omega_1$ и $\omega_2$, соответствующих пустому и заполненному контейнерам:

$$p_k = p(x_k = 1/\omega_1), \quad 1 - p_k = p(x_k = 0/\omega_1), \quad k = \overline{1,P},$$

$$q_k = p(x_k = 1/\omega_2), \quad 1 - q_k = p(x_k = 0/\omega_2), \quad k = \overline{1,P},$$

Тогда выражения для функций правдоподобия классов и логарифма отношения правдоподобия можно записать в виде

$$p(x/\omega_1) = \prod_{k=1}^{P} p_k^{x_k}(1-p_k)^{1-x_k},$$

$$p(x/\omega_2) = \prod_{k=1}^{P} q_k^{x_k}(1-q_k)^{1-x_k},$$

$$g(x) = \sum_{k=1}^{P} \ln \frac{p(x_k/\omega_1)}{p(x_k/\omega_2)} =$$

$$\sum_{k=1}^{P}\left( x_k \ln \frac{p_k}{q_k} + (1-x_k)\ln \frac{1-p_k}{1-q_k}\right) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} l_0 = \ln \frac{p(\omega_2)}{p(\omega_1)}.$$

Если предположить, что для всех признаков каждого класса вероятности единиц и нулей одинаковы $p_k = p \neq 0$, $q_k = q \neq 0$, $k = \overline{1,P}$, причем $p < q$. Учитывая тогда, что величина $\ln\left(p(1-q)/q(1-p)\right)$ меньше нуля, приведенное выше решающее правило преобразуется к виду

$$g(x) = L_x \ln \frac{p}{q} + M_x \ln \frac{1-p}{1-q} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} l_0,$$

$$L_x = \sum_{k=1}^{P} x_x, \quad M_x = \sum_{k=1}^{P}(1-x_x) = n - L_x, \qquad (1)$$

$$L_x = \sum_{k=1}^{P} x_k \underset{\omega_1}{\overset{\omega_2}{\gtrless}} L_0 = \left( l_0 - P\ln \frac{1-p}{1-q}\right) \Big/ \ln \frac{p(1-q)}{q(1-p)},$$

где $L_x, M_x$ — количество единиц (ответов «да») и количество нулей (ответов «нет»), полученные в ходе наблюдения.

Данная ситуация означает, что фактически проводится «опрос» $P$ независимых и равноценных признаков и сравнение общего количества полученных единиц (или нулей) с пороговым значением, зависящим от априорных вероятностей классов и соотношения вероятностей частных решений, непосредственно связанных с вероятностями ошибок первого и второго рода при первичной обработке $p \simeq er_{12}$, $1 - q \simeq er_{21}$, получаемых при тестировании нейросетевого классификатора, используемого для анализа блоков.

Такое представление алгоритма (1) позволяет, используя схему Бернулли для описания $P$ независимых испытаний, записать в данном случае точные выражения для вероятностей ошибок распознавания [31]. Однако в нашем случае трудно ожидать, что вероятности принятия частных решений в блоках бу-

дут одинаковы в пределах всего изображения большого размера. Это связано со спецификой используемой схемы встраивания на основе алгоритмов WOW и S-UNIWARD, которые заполняет контейнер в пространственном отношении неравномерно. Поэтому в наших исследованиях подбор оптимального порога $L_0$. как и оценка вероятностей правильных (ошибочных) решений, проводились экспериментально.

В качестве модификации алгоритма (1) также рассматривался адаптивный алгоритм, предусматривающий подсчет числа решений не по всем фрагментам (блокам) анализируемого в данный момент изображения, а только по тем из них, которые потенциально могут содержать определенное отличное от нуля число модифицированных пикселей. Для этого проводится анализ каждого блока с точки зрения его зашумленности и выбираются только те блоки, в которых число потенциально модифицированных пикселей может быть больше заданного порога. Простейшим способом подобного отбора блоков является встраивание в него произвольной псевдослучайной последовательности с помощью одного из алгоритмов WOW, HUGO, S-UNIWARD и подсчета числа измененных пикселей. Очевидно, что полученный в этом случае результат может быть близок к реальному числу потенциально модифицируемых пикселей в блоке. Таким образом, алгоритм (1) преобразуется к виду

$$L_x = \sum_{k=1}^{P_{ch}} x_{i_k} \begin{matrix} > \omega_2 \\ < \omega_1 \end{matrix} L_{0,ch}, \quad R_{i_k} > \rho m_{stego},\ k = \overline{1,P}, \quad (2)$$

$$m_{stego} = \frac{1}{P}\sum_{k=1}^{P} R_k,$$

где $R_{i_k}$ — количество потенциально модифицированных пикселей в блоке с индексом $i_k$, $k = \overline{1,P}$, $m_{stego}$ — среднее арифметическое потенциального заполнения блоков в данном изображении; $\rho$ — коэффициент, определяющий порог, по которому блоки отбираются для использования в процессе принятия решения. Экспериментально установлено, что значения $\rho$ должны устанавливаться в диапазоне 0.25…1. Подбор оптимального порога может проводиться в ходе эксперимента с учетом ошибок первого и второго рода, полученных на этапе обучения и тестирования нейронной сети, анализирующей блоки. Выявлено, что чем меньше может быть величина $pl$, определяющая потенциально возможную полезную нагрузку изображения при встраивании в него ССИ, тем больше должен быть коэффициент $\rho$. Можно ожидать, что эффект от применения адаптивного алгоритма (2) будет проявляться при использовании блоков малых размеров и анализе изображений с минимальной возможной полезной нагрузкой.

### 2.3. Методика обучения сверточной сети

В качестве исходного датасета при проведении исследований использовалась база данных PPG-LIRMM-COLOR, содержащая 10000 цветных изображений размером 512×512. Для создания и встраивания стегосообщений в эти изображения использовались алгоритмы адаптивной пространственной стеганографии WOW и S-UNIWARD в реализации симулятора, загруженного из [32]. В качестве стегосообщения при этом генерировалась псевдослучайная последовательность с оригинальным для каждого изображений установочным ключом. При заполнении контейнеров-изображений рассматривалось два варианта: заполнение одного канала цветности (в нашем случае синего) и заполнение с одинаковой полезной нагрузкой всех трех каналов цветности. Таким образом, всего использовалось 20000 изображений: 10000 исходных и 10000 заполненных стегосообщениями.

Для обучения предлагаемой нейронной сети использовалось 16000 изображений: 8000 исходных и 8000 заполненных стегосообщениями. Для валидации и тестирования использовалось 4000 оставшихся изображений: 2000 исходных и 2000 заполненных. При обучении и тестировании сети с различными размерами входного блока из исходных и заполненных изображений вырезались небольшие изображения соответствующего размера $n_x \times m_y$ со случайным смещением и формиро-

вались подвыборки из такого же числа изображений малого размера.

С учетом пространственной неравномерности заполнения контейнеров стегосообщениями, чтобы исключить использование при обучении пустых фрагментов заполненных изображений, проводился анализ каждого блока с точки зрения его зашумленности с помощью алгоритма WOW и подсчета потенциального числа измененных пикселей (как для исследования по WOW, так и по S-UNIWARD). При формировании датасета в каждом изображении (как исходном, так и заполненном) выбирались только те блоки, в которых число потенциально модифицированных пикселей было больше порога $R_{i_k} > 0,5m_{stego}$, $k = \overline{1,P}$. Затем из полученного списка блоков случайным образом выбирался один из этих блоков. Так формировалась обучающая и валидационная подвыборка из 16000 и 4000 изображений размера $n_x \times m_y$. Кроме того, из 4000 исходных полноразмерных изображений, не участвующих в формировании обучающей подвыборки, формировалась аналогичным образом тестирующая подвыборка со случайным выбором участвующих в ней блоков. Это обеспечивало практическую независимость тестирующей и валидационной подвыборок.

При обучении сети использовался оптимизатор adam на 30 эпохах с начальной скоростью 0.001, параметром L2-регуляризации 0.001, размером минибатча 64. Затем проводилось дообучение сети на 10 эпохах с начальной скоростью 0.0001, параметром L2-регуляризации 0.0001.

При обучении нейронных сетей на изображениях с малым уровнем полезной нагрузки может случиться ситуация, когда сеть из-за слишком малых различий пустых и заполненных изображений не выходит в режим обучения. Для преодоления этой ситуации целесообразно проводить задание стартовых значений весовых коэффициентов с использованием сети, ранее обученной на изображениях для более высоких уровней полезной нагрузки, и переобучать ее на изображениях с малым уровнем полезной нагрузки. Подобная стратегия оказалась весьма эффективной и в ситуациях, когда сети удавалось обучаться самостоятельно при малой полезной нагрузке изображений.

Все алгоритмы обработки информации реализованы в среде Matlab с использованием возможностей пакета Deep Learning Toolbox.

## 3. РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

В соответствии с представленной выше методикой проводилось обучение и тестирование предложенной сверточной нейронной сети, а также общего алгоритма вторичной обработки, обеспечивающих обнаружение факта скрытия стегосообщения в цветных изображениях. Внедрение стегосообщений проводилось с использованием алгоритмов WOW, S-UNIWARD при различных объемах полезной нагрузки $pl$ и заполнении каналов цветности. В табл. 3 представлены результаты, полученные при валидации/тестировании обученной сети на блоках изображений небольших размеров.

Анализ полученных результатов показывает, что при заполнении всех трех цветовых каналов точность обнаружения существенно повышается также, как она повышается и при увеличении размера анализируемого фрагмента. При малой нагрузке $pl = 0.2$ и только в одном цветовом канале, что соответствует реальному заполнению контейнера еще в 3 раза меньше, точность обработки не выше 60 %, при этом величина ошибки первого рода, когда пустой фрагмент принимается за заполненный, может превышать 50 % при существенно меньшей ошибке второго рода.

После обучения нейронных сетей с различными размерами входного фрагмента эти сети сохранялись для выполнения вторичной обработки совокупности блоков, формируемых в пределах изображения большого размера и последовательно подаваемых на вход сети. На рис. 3 показаны типичные гистограммы, описывающие распределение числа ошибок первого и второго рода после вторичной обработки и точности классификации изображений в зависимости от отношения порога $L_0$ к общему числу анализируемых блоков $P = 1024$ при размере блока 32×32. Внедрение стегосообщения здесь осуществлялось алго-

*Таблица 3. Результаты оценки точности обнаружения в блоках для предлагаемой архитектуры нейронной сети при различных объемах внедрения (полезной нагрузке) в % [Table 3. The results of evaluating the detection accuracy in blocks for the proposed architecture of the neural network for various embedding volumes (payload) in %]*

| Анализируемый алгоритм ССИ | WOW с загрузкой контейнера в одном цветовом канале | | WOW с загрузкой контейнера в трех цветовых каналах | | S-UNIWARD с загрузкой контейнера в одном цветовом канале | | S-UNIWARD с загрузкой контейнера в трех цветовых каналах | |
|---|---|---|---|---|---|---|---|---|
| Размеры входного изображения (блока) | $pl = 0.2$ | $pl = 0.4$ | $pl = 0.2$ | $pl = 0.4$ | $pl = 0.2$ | $pl = 0.4$ | $pl = 0.2$ | $pl = 0.4$ |
| $n_x = m_y = 32$ | 54.43/ 54.83 | 59.62/ 60.55 | 62.48/ 61.60 | 73.90/ 72.45 | 53.13/ 53.37 | 59.75/ 60.10 | 60.80/ 61.35 | 71.98/ 73.12 |
| $n_x = m_y = 64$ | 56.90/ 56.05 | 66.50/ 66.27 | 66.10/ 65.75 | 81.87/ 81.57 | 55.35/ 55.37 | 65.32/ 65.70 | 65.68/ 64.82 | 81.87/ 82.57 |
| $n_x = m_y = 128$ | 59.92/ 59.78 | 73.72/ 74.05 | 77.30/ 76.18 | 87.68/ 87.90 | 60.12/ 60.62 | 73.95/ 75.55 | 73.12/ 73.62 | 86.25/ 85.62 |

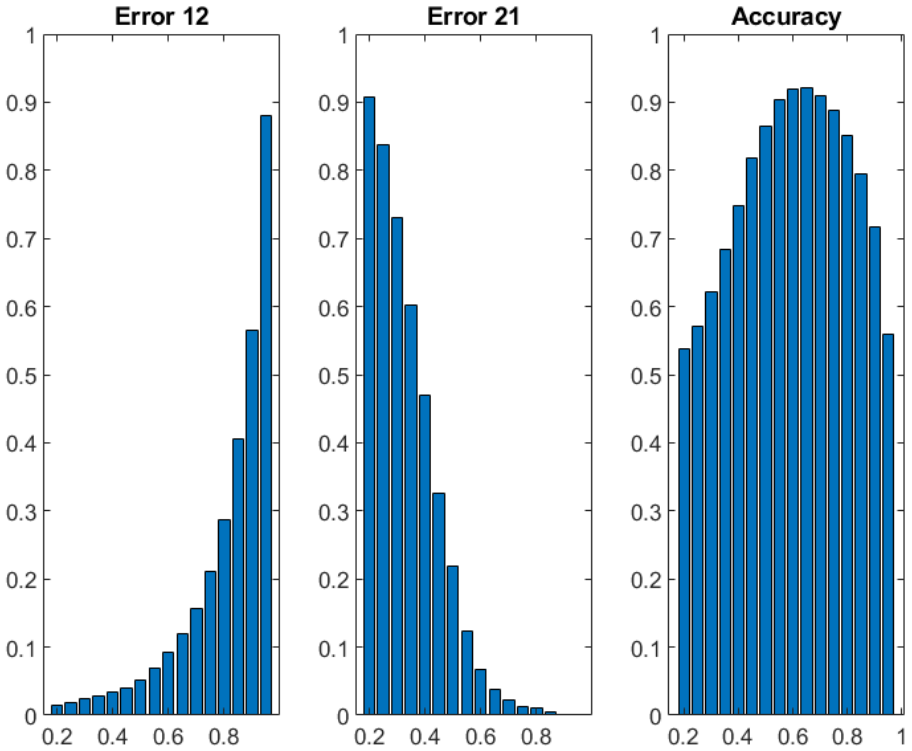


*Рис. 3. Типичные гистограммы распределения числа ошибок первого и второго рода и точности классификации изображений от отношения $L_0 / P$*
*[Fig. 3. Typical histograms of the first and second kind errors number distribution and the accuracy of image classification on the $L_0 / P$ ratio]*

ритмом WOW в трех каналах цветности при объеме полезной нагрузки $pl = 0.4$. Максимальное значение точности стегоанализа 92.10 достигается при отношении $L_0 / P = 0.65$. При использовании алгоритма (2) максимальное значение точности 92.85 достигается при отношении $L_{0,ch} / P_{ch} = 0.5$. В табл. 4 представлены результаты, полученные при тестирова-

*Таблица 4. Результаты оценки точности обнаружения по совокупности блоков для изображений большого размера при различных объемах внедрения (полезной нагрузке) в % [Table 4. The results of evaluating the detection accuracy by a set of blocks for large images at various embedding volumes (payload) in %]*

| Анализируемый алгоритм ССИ | WOW с загрузкой контейнера в одном цветовом канале | | WOW с загрузкой контейнера в трех цветовых каналах | | S-UNIWARD с загрузкой контейнера в одном цветовом канале | | S-UNIWARD с загрузкой контейнера в трех цветовых каналах | |
|---|---|---|---|---|---|---|---|---|
| Размеры используемых блоков | $pl = 0.2$ | $pl = 0.4$ | $pl = 0.2$ | $pl = 0.4$ | $pl = 0.2$ | $pl = 0.4$ | $pl = 0.2$ | $pl = 0.4$ |
| $n_x = m_y = 32$ $P = 1024$ | 55.97/ 59.02 | 70.38/ 71.50 | 81.30/ 82.53 | 92.10/ 92.85 | 56.40/ 57.95 | 65.63/ 67.30 | 77.90/ 79.03 | 91.47/ 91.73 |
| $n_x = m_y = 64$ $P = 256$ | 58.07/ 62.50 | 76.57/ 77.28 | 82.65/ 82.17 | 93.92/ 93.83 | 58.13/ 59.02 | 74.28/ 75.35 | 82.42/ 82.82 | 94.07/ 94.20 |
| $n_x = m_y = 128$ $P = 16$ | **65.38/ 66.32** | **82.57/ 82.28** | **87.30/ 88.00** | **94.72/ 94.47** | **66.30/ 67.35** | **81.37/ 81.88** | **84.42/ 84.45** | **94.45/ 94.45** |

нии двух алгоритмов обработки (1), (2) на изображениях размера 512×512 при различных размерах блоков и количестве блоков, подаваемых на вход при применении ранее обученной сверточной нейронной сети.

Анализ представленных результатов показывает, что предложенный подход и реализованные на его основе алгоритмы позволяют достичь точности обнаружения стегосообщений, сопоставимой с лучшими результатами, представленными в ранее опубликованных работах, а, в некоторых случаях, и превышающей их. Также следует отметить, что, как и ожидалось, применение адаптивного алгоритма вторичной обработки оправдано в большей степени при малых размерах анализируемых блоков и, соответственно, большом их количестве, а также при меньшей величине полезной нагрузки. Этот вариант алгоритма позволяет получить прирост точности от 0.5% до 3.5%. Одновременно следует отметить, что при значениях $pl = 0.4$ оптимальное значение порога достигается при отношении $L_0 / P$, пределах 0.45…0.65. А для значения полезной нагрузки $pl = 0.2$ рекомендуемая величина этого отношения смещается в верхнюю сторону в диапазон значений 0.55…0.75, что объясняется наличием большего числа «пустых» фрагментов в заполненном контейнере. Большая ошибка первого рода в результатах при $pl = 0.2$ для одного канала цветности также присутствует, но ее можно уменьшить за счет изменения порогового отношения $L_0 / P$.

## ЗАКЛЮЧЕНИЕ

На основе анализа известных публикаций в области стегоанализа с использованием методов и алгоритмов машинного обучения определены достижимые в настоящее время показатели качества обнаружения стегосообщений в черно-белых и цветных изображениях. Отмечены перспективы использования для решения задачи стегоанализа методов глубокого обучения с применением сверточных нейронных сетей различной архитектуры.

В развитие результатов известных исследований предложен подход, основанный на использовании сверточных сетей для последовательного анализа небольших фрагментов (блоков) исходных больших изображений с последующим объединением получаемых результатов классификации в виде совокупности бинарных признаков по схеме наивного байесовского классификатора. Предложена относительно простая архитектура сверточной сети, состоящей из трех сверточных и

трех полносвязных обучаемых слоев. В первом сверточном слое дополнительно реализован слой пространственной высокочастотной фильтрации с возможностью гибкой перестройки параметров фильтра и гауссовская функция активация с настраиваемым в процессе обучения параметром влияния. В качестве алгоритма вторичной обработки результатов классификации совокупности блоков в пределах одного изображения использован простой алгоритм сравнения общего числа «положительных» и «отрицательных» решений с экспериментально подбираемым порогом. Показано, что реализованные на основе такого подхода алгоритмы позволяют выявлять факт наличия стеганографически скрытой информации при использовании алгоритмов внедрения WOW и S-UNIWARD с точностью, не уступающей результатам, полученных в работах других авторов, а, в ряде случаев, и превосходящей их.

Как одно из преимуществ предлагаемого подхода, также следует отметить независимость реализуемой схемы обработки от размера анализируемого изображения в той ее части, где проводится обучение нейронных сетей, а также возможность быстрого переобучения нейронных сетей, ранее обученных на изображениях с высокой полезной нагрузкой, для обнаружения стегосообщений на изображениях с малой полезной нагрузкой.

## КОНФЛИКТ ИНТЕРЕСОВ

Авторы декларируют отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.

## СПИСОК ЛИТЕРАТУРЫ

1. *Шелухин, О. И.* Стеганография. Алгоритмы и программная реализация. / О. И. Шелухин, С. Д. Канаев; под ред. профессора О. И. Шелухина, – Горячая линия - Телеком, 2017. – 592 с.

2. *Грибунин, В. Г.* Цифровая стеганография / В. Г. Грибунин, И. Н. Оков, И. В. Туринцев. – Москва : Солон-Пресс, 2002. – 272 с.

3. *Конахович, Г. Ф.* Принципы стеганографического анализа / Г. Ф. Конахович, А. Ю. Пузыренко // Компьютерная стеганография. Теория и практика. – Москва : МК-Пресс, 2006. – 288 с.

4. *Czaplewski, Bartosz.* Current trends in the field of steganalysis and guidelines for constructions of new steganalysis schemes / Bartosz Czaplewski // Przegląd Telekomunikacyjny + Wiadomości Telekomunikacyjne. – 2017. – No 10. – P. 1121–1125. – DOI: 10.15199/59.2017.10.3.

5. *Валишин, М. Ф.* Повышение эффективности методов противодействия встраиванию скрытой информации в графические файлы / М. Ф. Валишин. – Ульяновск, 2015. – 106 с.

6. *Lyu, S.* Detecting messages using higher-order statistics and support vector machines [Электронный ресурс] / S. Lyu, H. Farid // Режим доступа: http://hackerzvoice.net/madchat/crypto/stegano/ih02.pdf.

7. *Lyu, S.* Steganalysis using color wavelet statistics and one-class support vector machines [Электронный ресурс] / S. Lyu, H. Farid // DOI: 10.1117/12.526012. Режим доступа: http://www.ists.dartmouth.edu/library/34.pdf.

8. *Pevny, T.* Steganalysis by subtractive pixel adjacency matrix / T. Pevny, P. Bas, J. Fridrich // IEEE Trans. Information Forensics and Security. – 2010. – V. 5, No 2. – P. 215–224. – DOI: 10.1109/TIFS.2010.2045842.

9. *Fridrich, J.* Rich models for steganalysis of digital images / J. Fridrich // IEEE Trans. Inform. Forensics and Security. – 2012. – V. 7, No 3. – P. 868–882. – DOI: 10.1109/TIFS.2012.2190402.

10. *Holub, V.* Random projections of residuals for digital image steganalysis / V. Holub, J. Fridrich // IEEE Trans. Inform. Forensics and Security. – 2013. – V. 8, No 12. – P. 1996–2006. – DOI: 10.1109/TIFS.2013.2286682.

11. *Bas, P.* "Break our steganographic system" the ins and outs of organizing BOSS / P. Bas, T. Filler, T. Pevn´y // LNCS. – 2011. – V. 6958. – P. 59–70. – DOI: 10.1007/978-3-642-24178-9_5.

12. PPG-LIRMM-COLOR database. – Режим доступа: http://www.lirmm.fr/~chaumont/PPG-LIRMM-COLOR.html.

13. *Pevn´y, T.* Using high-dimensional image models to perform highly undetectable steg-

anography / T. Pevn´y, P. Bas, T. Filler // LNCS. – 2010. – V. 6387. – P. 161–177.

14. *Holub, V.* Digital image steganography using universal distortion / V. Holub, J. Fridrich // Proc. 1st ACM Workshop on Inform. Hiding and Multimedia Security (IHMMSec). – Montpellier, France, ACM, 2013. – P. 59–68. – DOI: 10.1145/2482513.2482514.

15. *Holub, V.* Designing steganographic distortion using directional filters / V. Holub, J. Fridrich // Proc. 4th IEEE Intern. Workshop on Inform. Forensics and Security (WIFS). – Tenerife, Spain, IEEE, 2012. – P. 234–239. – DOI: 10.1109/WIFS.2012.6412655.

16. *Жилкин, М. Ю.* Метод выявления скрытой информации, базирующийся на сжатии данных / М. Ю. Жилкин, Н. А. Меленцова, Б. Я. Рябко // Вычислительные технологии. – 2007. – Т. 12. – С. 26–31.

17. *Жилкин, М. Ю.* Стегоанализ графических данных на основе методов сжатия / М. Ю. Жилкин // Вестник СибГУТИ. – 2008. – № 2. – С. 62–66.

18. *Монарев, В. А.* Эффективное обнаружение стеганографически скрытой информации посредством интегрального классификатора на основе сжатия данных / В. А. Монарев, А. И. Пестунов // Прикладная дискретная математика. – 2018. – № 40. – С. 59–71.

19. *Kodovsky, J.* Ensemble classifiers for steganalysis of digital media / J. Kodovsky, V. Holub, J. Fridrich // IEEE Trans. Inform. Forensics and Security. – 2010. – V. 7, No 2. – P. 434–444. – DOI: 10.1109/TIFS.2011.2175919.

20. *Goljan, M.* Rich model for steganalysis of color images / M. Goljan, J. Fridrich, R. Cogranne // In IEEE Workshop on Information Forensic and Security, GA – 2014. – DOI: 10.1109/WIFS.2014.7084325.

21. *Abdulrahman, H.* Color image steganalysis using correlations between RGB channels / H. Abdulrahman, M. Chaumont, P. Montesinos, B. Magnier // Availability Reliability and Security (ARES), 10th International Conference on IEEE. – 2015. – P. 448–454. – DOI: 10.1109/ARES.2015.44.

22. *Mudhafar, M.* Steganalysis of Color Images for Low Payload Detection / M. Al-Jarrah Mudhafar, M. Al-Manaseer Renad // IHIP 2019: Proceedings of the 2019 2nd International Conference on Information Hiding and Image Processing. – September 2019. – P. 35–38. Режим доступа: https://doi.org/10.1145/3383913.3383915.

23. *Tabares-Soto, R.* Deep Learning Applied to Steganalysis of Digital Images: A Systematic Review / Reinel Tabares-Soto, Raúl Ramos-Pollán, Gustavo Isaza // Computer Science IEEE Access. – 2019. – DOI:10.1109/ACCESS.2019.2918086.

24. *Qian, Y.* Deep learning for steganalysis via convolutional neural networks / Y. Qian, J. Dong, W. Wang, T. Tan // Proceedings Volume 9409, Media Watermarking, Security, and Forensics 2015 / Ed. by A. M. Alattar, N. D. Memon, C. D. Heitzenrater. – SPIE, 2015, SPIE/IS&T Electronic Imaging. – San Francisco, California, United State, 2015. – DOI: 10.1117/12.2083479.

25. *Qian, Y.* Feature learning for steganalysis using convolutional neural networks / Y. Qian, J. Dong, W. Wang, T. Tan // Multimedia Tools and Applications. 2017. – V. 77, No 15. – P. 19633–19657.

26. *Couchot, J.-F.* Steganalysis via a Convolutional Neural Network using Large Convolution Filters [Электронный ресурс] / J.-F. Couchot, R. Couturier, C. Guyeux, M. Salomon // CoRR. – 2016. – Режим доступа: http://arxiv.org/abs/1605.07946.

27. *Pibre, L.* Deep learning is a good steganalysis tool when embedding key is reused for different images, even if there is a cover source-mismatch / Lionel Pibre, Pasquet Jerome, Dino Ienco, Marc Chaumont // Conference: Media Watermarking, Security, and Forensics, Part of IS&T International Symposium on Electronic Imaging, EI'2016. – San Francisco, California, USA, 2016.

28. *Нагорный, Н. А.* Исследование алгоритмов стегоанализа изображений с использованием глубоких нейронных сетей / Н. А. Нагорный А. А. Сирота // Сборник студенческих научных работ факультета компьютерных наук ВГУ. – 2019. – Часть 2. – С.145–151.

29. *Yedroudj, M.* Yedrouj-Net: An Efficient CNN for Spatial Steganalysis / M. Yedroudj, F. Comby, M. Chaumont // In Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing. – 2018. – P. 2092–2096.

30. *Полунин, А. А.* Использование аппарата сверточных нейронных сетей для стегоанали-

за цифровых изображений / А. А. Полунин, Э. А. Яндашевская // Труды ИСП РАН. – 2020. – Том 32, вып. 4. – С. 155–164. – DOI: 10.15514/ISPRAS-2020-32(4)-11.

31. *Сирота, А. А.* Методы и алгоритмы анализа данных и их моделирование в MATLAB: [учебное пособие] / А. А. Сирота. – Санкт-Петербург: БХВ- Петербург, 2016. – 381 с.

32. Digital Data Embedding Laboratory [Электронный ресурс] / Department of Electrical and Computer Engineering SUNY Binghamton, Binghamton, NY 13902-6000. – Режим доступа: http://dde.binghamton.edu/download/stego_algorithms.

**Сирота А. А.** — д-р техн. наук, проф., заведующий кафедрой технологий обработки и защиты информации, факультет компьютерных наук, Воронежский государственный университет.
E-mail: sir@cs.vsu.ru
ORCID iD: https://orcid.org/0000-0002-5785-8513

**Дрюченко М. А.** — канд. техн. наук, доц., доцент кафедры технологий обработки и защиты информации, факультет компьютерных наук, Воронежский государственный университет.
E-mail: m_dryuchenko@mail.ru
ORCID iD: https://orcid.org/0000-0001-8837-5875

**Иванков А. Ю.** – канд. физ.-мат. наук, доцент кафедры технологий обработки и защиты информации, факультет компьютерных наук, Воронежский государственный университет.
E-mail: ivankov@cs.vsu.ru
ORCID iD: https://orcid.org/0000-0002-3017-6037

DOI: https://doi.org/10.17308/sait.2021.1/3369 ISSN 1995-5499
Received 19.03.2021
Accepted 26.04.2021

# STEGANALYSIS OF DIGITAL IMAGES BY MEANS OF SHALLOW AND DEEP MACHINE LEARNING: EXISTING APPROACHES AND NEW SOLUTIONS

**© 2021 A. A. Sirota, M. A. Dryuchenko, A. Y. Ivankov**✉

*Voronezh State University*
*1, Universitetskaya Square, 394018 Voronezh, Russian Federation*

**Annotation.** The article considers the current state of the problem of steganalysis of digital images in order to develop and study effective methods of revealing hidden (invisible) messages in container images. In the first part of the article, we provide a classification of the existing approaches and detail the previously obtained results of steganalysis performed using shallow and deep machine learning methods. We also describe the indicator systems used in shallow machine learning today and classifiers based on them (ensemble methods, support vector machines, etc.). An alternative method is based on using deep convolutional neural networks with various modifications (additional layers, special activation functions, etc.). The paper presents the results of the comparative analysis of the effectiveness of different approaches and different neural network architectures used in steganalysis. The analysis was

✉ Ivankov Alexander Y.
e-mail: ivankov@cs.vsu.ru

performed using standard image sets. Hidden messages were embedded using adaptive spatial steganography algorithms: WOW, HUGO, and S-UNIWARD. The study demonstrated the universality and effectiveness of deep machine learning and showed that it is a promising method that can be used in steganalysis. In the second part of the article, we suggest a new architecture for a deep neural network and describe its performance when applied in the steganalysis of colour images. The key idea of the suggested approach is to use relatively simple convolutional networks for subsequent analysis of small fragments (blocks) of initial large images. The obtained classification results are then fused in a sequence of binary features using a Naive Bayes classifier. The experiments were performed using the PPG-LIRMM-COLOR database. WOW and S-UNIWARD algorithms were used to embed steganographic messages with various payload. The precision of the steganalysis of large images is compatible with and, in some cases, even better than the results obtained by other authors.

**Keywords:** steganography, steganalysis, stegmessage, digital images, machine learning, deep neural networks.

## CONFLICT OF INTEREST

The authors declare the absence of obvious and potential conflicts of interest related to the publication of this article.

## REFERENCES

1. *Sheluhin O. I. & Kanaev S. D.* (2017) Steganografiya. Algoritmy i programmnaya realizatsiya [Steganography. Algorithms and software implementation], In O. I. Sheluhina (ed.). Moscow, Goryachaya liniya - Telekom. (in Russian)

2. *Gribunin V. G., Okov I. N. & Turintsev I. V.* (2002) Tsifrovaya steganografiya [Digital Steganography]. Moscow, Solon-Press. (in Russian)

3. *Konahovich G. F. & Puzyrenko A. Ju.* (2006) Komp'juternaja steganografija. Teorija i praktika [Computer-held steganography. Theory and practice] K.: MK-Press. (In Russia)

4. *Czaplewski Bartosz* (2017) Current trends in the field of steganalysis and guidelines for constructions of new steganalysis schemes. Przegląd Telekomunikacyjny + Wiadomości Telekomunikacyjne. (10), 1121–1125. Available from: doi:10.15199/59.2017.10.3.

5. *Valishin M. F.* (2015) Povyshenie effectivnosti metodov protivodeystviya vstraivaniyu skritoy informatsii v graficheskie faily [Improving the effectiveness of methods to counter the embedding of hidden information in graphic files]. Ulyanovsk. (in Russian)

6. *Lyu S. & Farid H.* Detecting messages using higher-order statistics and support vector machines [Electronic resource]. Available at: http://hackerzvoice.net/madchat/crypto/stegano/ih02.pdf.

7. *Lyu S. & Farid H.* (2004) Steganalysis using color wavelet statistics and one-class support vector machines [Electronic resource]. Available from: doi:10.1117/12.526012. Available at: http://www.ists.dartmouth.edu/library/34.pdf.

8. *Pevny T., Bas P. & Fridrich J.* (2010) Steganalysis by subtractive pixel adjacency matrix. IEEE Trans. Information Forensics and Security. 5 (2), 215–224. Available from: doi:10.1109/TIFS.2010.2045842.

9. *Fridrich J.* (2012) Rich models for steganalysis of digital images. IEEE Trans. Inform. Forensics and Security. 7 (3), 868–882. Available from: doi:10.1109/TIFS.2012.2190402.

10. *Holub V. & Fridrich J.* (2013) Random projections of residuals for digital image steganalysis / V. Holub, J. Fridrich // IEEE Trans. Inform. Forensics and Security. 8 (12), 1996–2006. Available from: doi:10.1109/TIFS.2013.2286682.

11. *Bas P., Filler T. & Pevn´y T.* (2011) "Break our steganographic system" the ins and outs of organizing BOSS. LNCS. 6958, 59–70. Available from: doi:10.1007/978-3-642-24178-9_5.

12. PPG-LIRMM-COLOR database. Available at: http://www.lirmm.fr/~chaumont/PPG-LIRMM-COLOR.html.

13. *Pevn´y T., Bas P. & Filler T.* (2010) Using high-dimensional image models to perform

highly undetectable steganography. LNCS. 6387, 161–177.

14. *Holub V. & Fridrich J.* (2013) Digital image steganography using universal distortion. Proc. 1st ACM Workshop on Inform. Hiding and Multimedia Security (IHMMSec), 2013, Montpellier, France, ACM, pp. 59–68. Available from: doi:10.1145/2482513.2482514.

15. *Holub V. & Fridrich J.* (2012) Designing steganographic distortion using directional filters. Proc. 4th IEEE Intern. Workshop on Inform. Forensics and Security (WIFS), 2012, Tenerife, Spain, IEEE, pp. 234–239. Available from: doi:10.1109/WIFS.2012.6412655.

16. *Zhilkin M. U., Melentsova N. A. & Ryabko B. Ja.* (2007) Metod vyjavlenija skrytoy informatsii, bazirujuschijsya na szhatii dannyh [Method for revealing hidden information based on data compression]. Journal of Computational Technologies. (12), 26-31. (in Russian)

17. *Zhilkin, M. U.* (2008) Stegoanaliz graficheskih dannyh na osnove metodov szhatiya [Steganalysis of graphic data based on compression methods]. Vestnik SibGUTI. (2), 62–66. (in Russian)

18. *Monarev V. A. & Pestunov A. I.* (2018) Efficient steganography detection by means of compression-based integral classifier. Prikladnaya Diskretnaya Matematika (40) 59–71. Available from: doi:10.17223/20710410/40/5.

19. *Kodovsky J., Holub V. & Fridrich J.* (2010) Ensemble classifiers for steganalysis of digital media. IEEE Trans. Inform. Forensics and Security. 7 (2), 434–444. Available from: doi:10.1109/TIFS.2011.2175919.

20. *Goljan M., Fridrich J. & Cogranne R.* (2014) Rich model for steganalysis of color images. In IEEE Workshop on Information Forensic and Security, 2014, GA. — 2014. Available from: doi:10.1109/WIFS.2014.7084325.

21. *Abdulrahman H., Chaumont M., Montesinos P. & Magnier B.* (2015) Color image steganalysis using correlations between RGB channels. Availability Reliability and Security (ARES), 10th International Conference on IEEE, pp. 448-454. Available from: doi:10.1109/ARES.2015.44.

22. *Mudhafar M. & Renad M.* (2019) Steganalysis of Color Images for Low Payload Detection. IHIP 2019: Proceedings of the 2019 2nd International Conference on Information Hiding and Image Processing, September 2019, pp 35–38. Available from: doi:10.1145/3383913.3383915.

23. *Tabares-Soto R. & Ramos-Pollán R.* (2019) Deep Learning Applied to Steganalysis of Digital Images: A Systematic Review. Computer Science IEEE Access. Available from: doi:10.1109/ACCESS.2019.2918086.

24. *Qian Y., Dong J., Wang W. & Tan T.* (2015) Deep learning for steganalysis via convolutional neural networks. Proceedings Volume 9409, Media Watermarking, Security, and Forensics, 2015, San Francisco, California, United State. Available from: doi:10.1117/12.2083479.

25. *Qian Y., Dong J., Wang W. & Tan T.* (2017) Feature learning for steganalysis using convolutional neural networks. Multimedia Tools and Applications. 77 (15), 19633–19657.

26. *Couchot J.-F., Couturier R., Guyeux C. & Salomon M.* (2016) Steganalysis via a Convolutional Neural Network using Large Convolution Filters [Electronic resource]. CoRR. Available at: http://arxiv.org/abs/1605.07946.

27. *Pibre L., Jerome P., Ienco D. & Chaumont M.* (2016) Deep learning is a good steganalysis tool when embedding key is reused for different images, even if there is a cover source-mismatch. Conference: Media Watermarking, Security, and Forensics, Part of IS&T International Symposium on Electronic Imaging, EI'2016, 2016, San Francisco, California, USA.

28. *Nagorny N. A. & Sirota A. A.* (2019) Issledovanie algoritmov stegoanaliza izobrazheniy s ispol'zovaniem glubokih neironnyh setey [Investigation of image steganalysis algorithms using deep neural networks]. Sbornik studencheskih nauchnyh rabot faculteta comp'uternyh nauk VGU. (2) 145–151. (in Russian)

29. *Yedroudj, M., Comby, F. & Chaumont M.* Yedrouj-Net: An Efficient CNN for Spatial Steganalysis. In Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing, 2018, P. 2092–2096.

30. *Polunin, A. A. & Yandashevskaya E. A.* (2020) Using of convolutional neural networks for steganalysis of digital images. Proceedings of the Institute for System Programming of the RAS 32 (4) 155–164. Available from: doi:10.15514/ISPRAS-2020-32(4)-11.

31. *Sirota, A. A.* (2016) Metody I algoritmy analiza dannyh i ih modelirovanie v MATLAB [Methods and algorithms for data analysis and modeling in MATLAB]. St. Petersburg: BHV-Peterburg. (in Russian)

32. Digital Data Embedding Laboratory [Electronic resource]. Department of Electrical and Computer Engineering SUNY Binghamton, Binghamton, NY 13902-6000. Available at: http://dde.binghamton.edu/download/stego_algorithms.

**Sirota Alexander A.** — DSc in Technical Sciences, Head of the Department of Information Security and Processing Technologies, Faculty of Computer Sciences, Voronezh State University.
E-mail: sir@cs.vsu.ru
ORCID iD: https://orcid.org/0000-0002-5785-8513

**Dryuchenko Mikhail A.** — PhD in Technical Sciences, Associate Professor, Department of Information Security and Processing Technologies, Faculty of Computer Sciences, Voronezh State University.
E-mail: m_dryuchenko@mail.ru
ORCID iD: https://orcid.org/0000-0001-8837-5875

**Ivankov Alexander Y.** – PhD in Technical Sciences, Associate Professor, Department of Information Security and Processing Technologies, Faculty of Computer Sciences, Voronezh State University.
E-mail: ivankov@cs.vsu.ru
ORCID iD: https://orcid.org/0000-0002-3017-6037