DOI: 10.15514/ISPRAS–2020–32(4)–11

Использование аппарата свёрточных нейронных сетей для стегоанализа цифровых изображений

А.А. Полунин, ORCID: 0000-0002-5870-5439 <polunin2002@mail.ru>
Э.А. Яндашевская, ORCID: 0000-0003-1050-9137 <elenayanda@yandex.ru>
Академия Федеральной службы охраны Российской Федерации,
302015, Россия, г. Орел, ул. Приборостроительная, д. 35

Аннотация. В статье дается обоснование актуальности задачи стегоанализа, как определения факта наличия скрытого канала в инфокоммуникационных системах, узлы которых обмениваются цифровыми изображениями. Рассматриваются вопросы применения аппарата свёрточных нейронных сетей для решения этой задачи. Предполагается, что вероятность правильной классификации изображений с помощью хорошо обученной свёрточной нейронной сети будет сопоставима с показателями статистических алгоритмов или RM-модели или даже окажется лучше них. Дается представление о принципах построения и возможностях свёрточных нейронных сетей в рамках их применимости к решению задачи стегоанализа. Для повышения оперативности и результативности процесса распознавания стегоконтейнеров предложен вариант модели классификации изображений для свёрточной нейронной сети, в которой используется комбинация нескольких свёрточных и полносвязных слоев. Разработана программная реализация варианта этой модели с возможностью обучения нейронной сети и оценивания качества классификации. Проведен анализ существующих программных продуктов, предназначенных для задачи определения факта использования стеганографии в цифровых изображениях. Обосновано преимущество классификаторов на основе нейронных сетей по сравнению со статистическими классификаторами. С использованием разработанной программной реализации проведено экспериментальное исследование модели классификации на наборах цифровых изображений, содержащихся в открытых источниках. В статье приведены результаты обучения нейронной сети, а также анализ сильных и слабых сторон выбранной модели.

Ключевые слова: стеганография; стегоанализ; цифровые изображения; нейронная сеть; свёрточный слой; полносвязный слой; машинное обучение

Для цитирования: Полунин А.А., Яндашевская Э.А. Использование аппарата свёрточных нейронных сетей для стегоанализа цифровых изображений. Труды ИСП РАН, том 32, вып. 4, 2020 г., стр. 155–164. DOI: 10.15514/ISPRAS–2020–32(4)–11

Using of convolutional neural networks for steganalysis of digital images

A.A. Polunin, ORCID: 0000-0002-5870-5439 <polunin2002@mail.ru>
E.A. Yandashevskaya, ORCID: 0000-0003-1050-9137 <elenayanda@yandex.ru>
Russian Federation Security Guard Service Federal Academy,
25, Priborostroitelnaya st., Oryol, 302015, Russia

Abstract. The article substantiates the relevance of steganalysis, as a determination of the presence of a hidden channel in telecommunication systems, whose nodes exchange digital images. The article deals with the application of convolutional neural networks to solve this problem. It is assumed that the probability of correct

image classification using a well-trained convolutional neural network will be comparable or even better than characteristics of statistical algorithms or the RM model. We introduce principles of construction and capabilities of convolutional neural networks in the framework of their applicability to solving the problem of steganalysis. To improve the efficiency and effectiveness of the stegocontainer recognition process, a version of the image classification model for a convolutional neural network is proposed. It is based on combination of several convolutional and fully connected layers. We have developed software for this model version with the ability to train a neural network and evaluate the quality of classification. The analysis of existing software products designed for the task of determining the fact of using steganography in digital images is carried out. The advantage of classifiers based on neural networks in comparison with statistical ones is proved. Using the developed software, an experimental study of classification model on sets of digital images contained in open sources has been carried out. The article presents the results of neural network training, as well as an analysis of the strengths and weaknesses of the selected model.

Keywords: steganography; steganalysis; digital images; neural network; convolutional layer; fully connected layer; machine learning

For citation: Polunin A.A., Yandashevskaya E.A. Using of convolutional neural networks for steganalysis of digital images. Trudy ISP RAN/Proc. ISP RAS, vol. 32, issue 4, 2020. pp. 155–164 (in Russian). DOI: 10.15514/ISPRAS–2020–32(4)–11

1. Введение

Одним из актуальных направлений в предметных областях безопасности и мониторинга инфокоммуникационных систем являются вопросы формирования и эксплуатации скрытых каналов на основе стеганографических методов преобразования информации. Такие каналы используют функциональные возможности этих систем с целью порождения несанкционированных информационных потоков или потоков удаленного управления сервисами. Одним из наиболее распространенных видов стеганографических контейнеров (далее – стегоконтейнеров), применяемых для организации подобных скрытых каналов, являются цифровые изображения. В связи с этим являются актуальными задачи стегоанализа, формулируемые, как определение факта наличия скрытого канала в инфокоммуникационных системах, узлы которых обмениваются цифровыми изображениями.

В настоящее время большинство средств стегоанализа, предназначенных для обнаружения факта стеговложения, базируется на статистических методах [1], которые основываются на формировании модели изображения, позволяющей определить вектор признаков наличия стеговложения. К таким моделям относится, например, RM-модель (Rich Model), применяемая для обучения статистических классификаторов [2].

Их существенными недостатками являются: требовательность аналитических моделей, на которых основываются методы стегоанализа, к ряду параметров анализируемых контейнеров, что влияет на чувствительность моделей, и необходимость реализации полученных решений численными методами, что приводит к снижению оперативности и результативности процесса стегоанализа. В последнее время ряд исследований в области стегоанализа, в частности, цифровых изображений посвящен применению методов машинного обучения, основанных на применении искусственных нейронных сетей [1, 3, 4]. С использованием этих методов решается задача бинарной классификации цифровых изображений – разделения их множества на подмножества, содержащие и не содержащие стеговложения.

В рамках статьи рассматривается модель классификации, основанная на искусственной нейронной сети, которая содержит комбинацию свёрточных и полносвязных слоев, с целью определения значений показателя точности процесса классификации.

2. Особенности применения аппарата свёрточных нейронных сетей для решения задачи стегоанализа цифровых изображений

Аппарат свёрточных нейронных сетей (СНС) находит широкое применение при решении задачи обнаружения пространственных зависимостей в цифровых изображениях, а особенности формирования СНС позволяют уменьшить количество параметров и улучшить качество определения признаков [5].

В общем случае СНС состоят из следующих базовых блоков:

- свёрточные слои;
- слои подвыборки (пулинга);
- полносвязные слои.

Как правило, первый свёрточный слой отвечает за распознавание низкоуровневых признаков, а последующие слои объединяют их, переходя к более высокоуровневым признакам. Исследование возможностей СНС показывает, что на достаточно большом количестве цифровых изображений процесс обучения СНС позволяет оптимально настроить значения весов свёрточных слоев, необходимых для преобразования цифровых изображения, в пригодный для вычислительной системы вектор признаков, обеспечивающий повышение результативности процесса распознавания при минимизации его вычислительной сложности. Кроме того, особенностью СНС является возможность обнаружения каких-либо характеристик не в целом цифровом изображении, а во многих его частях, что достигается сегментацией изображения во время прохождения ядра СНС.

Функцией свёрточного слоя СНС является двумерная свертка – операция, используемая для уменьшения размера матрицы. Ядро K – матрица, которая состоит из коэффициентов, называемых весами, по сути является фильтром. При перемещении ядра по двумерному изображению I – другой матрице, выполняется умножение весов на значения пикселей, над которыми находится ядро, с их последующим суммированием, результатом которого является значение нового элемента (пикселя) следующего слоя (рис. 1). Таким образом, на выходе свёрточного слоя получается новая матрица (изображение) меньшего размера [6].

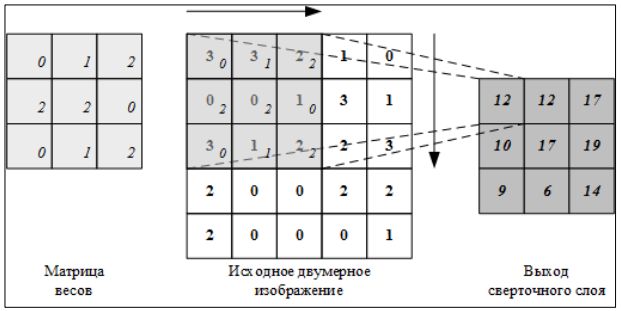

Рис. 1. Процесс двумерной свертки черно-белого цифрового изображения
Fig. 1. Two-dimensional convolution of a black-and-white digital image

Как и свёрточный, слой подвыборки (пулинговый) необходим для уменьшения размера изображения с целью минимизации количества требуемых вычислительных операций. В настоящее время используется несколько типов пулинга: максимальный (Max Pooling), средний (Average Pooling) и пулинг суммы (Sum Pooling). Первый тип применяется для вычисления максимального значения из части изображения, покрываемой ядром, второй – для вычисления среднего среди всех значений анализируемой области, третий – для определения их суммы (рис. 2).

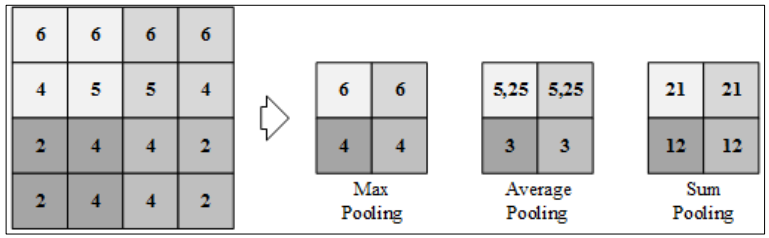

Рис. 2. Процесс обработки изображения слоем подвыборки
Fig. 2. Image processing with a pooling layer

Основной задачей полносвязного слоя (рис. 3) является моделирование нелинейной функции, используемой непосредственно для классификации, в то время как предшествующие слои являются средствами предобработки цифрового изображения. Применение нескольких слоев обученной СНС позволяет находить закономерности во входных данных и, анализируя их, формировать результат (одно или несколько решений), являющийся решением задачи. После обучения СНС на её выходе (в случае классификации) можно будет установить вероятность принадлежности изображения к тому или иному классу, что актуально для решения задачи распознавания стегоконтейнеров [8].

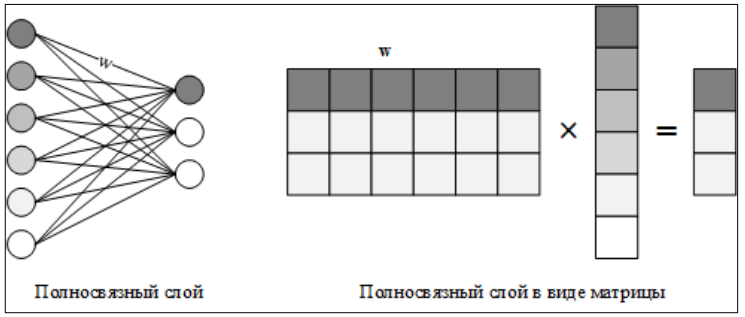

Рис. 3. Представление полносвязного слоя нейронной сети
Fig. 3. Neural network fully connected layer representation

3. Разработка модели свёрточной нейронной сети

Для синтеза свёрточного слоя СНС необходимо определить следующие параметры [9]:

- f (filters count) – количество фильтров в слое.
- K (kernel size) – размер (высота и ширина) ядра (обычно является нечётным числом, часто используются фильтры размером 3 или 5);
- s (stride) – шаг свёртки (количество пикселей, на которое перемещается матрица фильтра по входному изображению);
- p (padding) – дополнения нулями (количество пикселей, которые добавляются с каждого края изображения).

Таким образом, входными параметрами свёрточного слоя являются:

- собственно входное изображение в виде тензора $I_{in}(W_{in}, H_{in}, D_{in})$, где W_{in}, H_{in} – ширина и высота изображения соответственно, D_{in} – количество каналов;

- гиперпараметры: f, K, s, p.

Выходными данными слоя является тензор $I_{out}(W_{out}, H_{out}, D_{out})$, где:

$$W_{out} = \frac{W_{in} - K + 2p}{s} + 1, \quad (1)$$

$$H_{out} = \frac{H_{in} - K + 2p}{s} + 1, \quad (2)$$

$$D_{out} = f. \quad (3)$$

Слою подвыборки требуется всего один гиперпараметр – шаг пулинга, т.е. число раз, в которое нужно сократить пространственные размерности. Обычно используется слой пулинга с уменьшением размера входного тензора в два раза. Единственным гиперпараметром для полносвязного слоя является количество выходных значений.

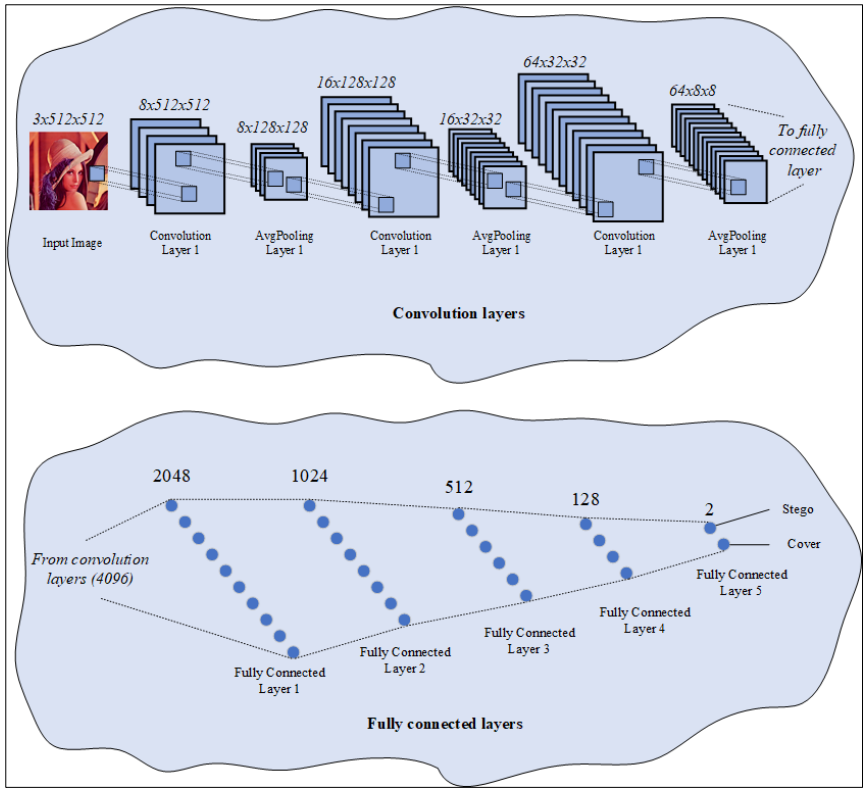

Рис. 4. Модель свёрточной нейронной сети
Fig. 4. Convolutional neural network model

Для решения стоящей в работе задачи стегоанализа цифровых изображений, была разработана модель СНС (рис. 4) со следующими слоями и их параметрами, рассчитанными в соответствии с выражениями (1–3):

- сверточный слой №1 (in_channels=3, out_channels=8, kernel_size=5, padding=2, stride=1);
- слой подвыборки №1(kernel_size=5, padding=2, stride=4);
- сверточный слой №2 (in_channels=8, out_channels=16, kernel_size=5, padding=2, stride=1);

- слой подвыборки №2 (kernel_size=5, padding=2, stride=4);
- сверточный слой №3 (in_channels=16, out_channels=64, kernel_size=5, padding=2, stride=1);
- слой подвыборки №3 (kernel_size=5, padding=2, stride=4);
- полносвязный слой №1 (in_features=4096, out_features=2048);
- полносвязный слой №2 (in_features=2048, out_features=1024);
- полносвязный слой №3 (in_features=1024, out_features=512);
- полносвязный слой №4 (in_features=512, out_features=128);
- полносвязный слой №4 (in_features=128, out_features=2).

После полной обработки входного цифрового изображения на основании выходных данных последнего слоя модели вычисляется вероятность принадлежности объекта к тому или иному классу.

4. Программная реализация модели свёрточной нейронной сети и ее экспериментальные исследования

Разработанная модель СНС была реализована в виде специального программного обеспечения (СПО) обнаружения стеганографических вложений в цифровых изображениях. Для решения задачи бинарной классификации были введены два класса цифровых изображений: «stego» и «cover». К первому относятся цифровые изображения со вложенной информацией – стегоконтейнеры, ко второму классу относятся цифровые изображения без вложений – пустые контейнеры. Тестирование СПО производилось на ресурсе Google Colaboratory. Данный сервис позволяет бесплатно использовать вычислительные мощности на удаленных серверах и быстро импортировать различные наборы данных с других ресурсов. СПО написано на языке Python, при этом для построения СНС использовалась библиотека с открытым исходным кодом PyTorch.

5. Результаты исследования

Наборы цифровых изображений для обучения СНС, реализованной в СПО, были взяты с ресурса kaggle.com. В качестве тренировочного набора данных были взяты изображения из digital-steganography (с применением стеганографических алгоритмов LSB, FFT, DCT – 15,7 тысячи изображений) и image-dataset (обычные изображения – 21,3 тысячи изображений). На входе в нейронную сеть каждое изображение масштабировалось к размеру 512x512. Обучение производилось по наборам из 64 изображений, после каждого набора высчитывалась функция потерь (loss) и точность (accuracy), их динамика в процессе обучения представлена на рис. 5.

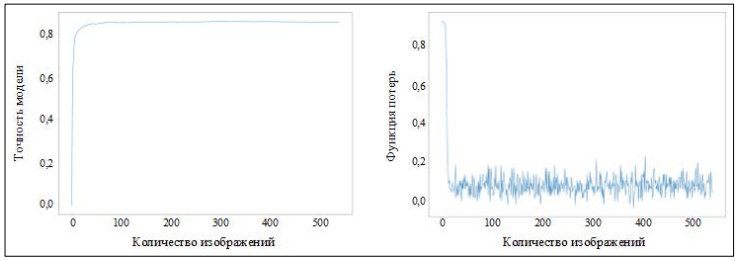

Рис. 5. Точность и функция потерь модели свёрточной нейронной сети в процессе обучения
Fig. 5. Convolutional neural network model training (accuracy and loss function)

Тестирование СПО производилось на наборе данных steghide-images, содержащем по 1,4 тысячи обычных изображений и содержащих стеганографические вложения. Точность и

функция потерь также определялись для каждого набора из 64 изображений. Их динамика приведена на рис. 6.

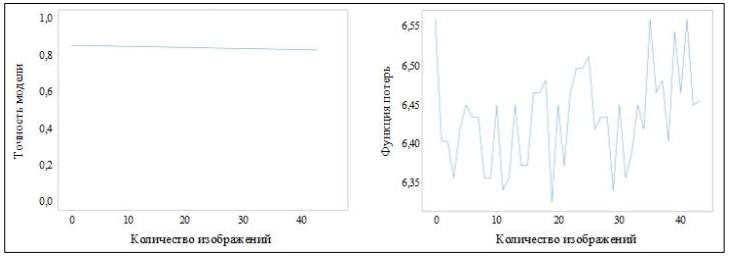

Рис. 6. Точность и функция потерь модели свёрточной нейронной сети в процессе тестирования
Fig. 5. Convolutional neural network model testing (accuracy and loss function)

Анализ полученных зависимостей (рис. 5-6) демонстрирует возможность правильной классификации в 85% случаев.

В [10] приведены результаты тестирования статистических классификаторов на изображениях, сформированных с использованием различных стеганографических алгоритмов (табл. 1). Из таблицы видно, что точность разработанной модели в целом превосходит средние показатели точности статистических классификаторов. Дополнительно следует отметить, что существенным недостатком статистических классификаторов, отсутствующим в методах на основе нейронных сетей, является их узкая специализация на строго определенные методы формирования стегоконтейнеров. Так в [3] представлены результаты анализа использования свёрточных нейронных сетей для решения задачи стегоанализа для алгоритмов встраивания WOW и S-UNIWARD при отношении количества вложенной информации на пиксель 0,2 bpp и 0,4 bpp (табл. 2).

Табл. 1. Сравнение вероятности ошибок при тестировании статистических классификаторов на различных стеганографических алгоритмах (вероятность ошибки)
Table 1. Error probability comparison during statistical classifiers testing with various steganographic algorithms

Алго-ритм	bpp	CHEN	CC-CHEN	LIU	CC-PEV	CDF	CC-300	CF	JRM	CC-JRM	J+ SRM
nsF5	0,05	0,4153	0,3816	0,3377	0,369	0,3594	0,3722	0,3377	0,3407	0,3298	0,3146
	0,1	0,3097	0,247	0,1732	0,2239	0,202	0,2207	0,1737	0,1782	0,1616	0,1375
	0,15	0,2094	0,1393	0,0706	0,1171	0,0906	0,1127	0,072	0,0793	0,0663	0,0468
	0,2	0,1345	0,0708	0,0273	0,0549	0,036	0,0486	0,0273	0,0338	0,0255	0,015
MBS	0,01	0,407	0,3962	0,3826	0,3876	0,3786	0,4038	0,371	0,3478	0,3414	0,326
	0,02	0,3178	0,2962	0,278	0,2827	0,2684	0,312	0,256	0,2156	0,2122	0,1832
	0,03	0,2395	0,21	0,1925	0,1965	0,1795	0,2241	0,1684	0,1266	0,1195	0,0983
	0,04	0,177	0,1437	0,1288	0,1298	0,1135	0,1594	0,1087	0,0751	0,067	0,0494
	0,05	0,1243	0,0946	0,0812	0,0833	0,0704	0,1176	0,0684	0,0427	0,0373	0,0282
YASS	0,077	0,2009	0,1825	0,2324	0,2279	0,1268	0,093	0,0532	0,0324	0,0303	0,0173
	0,114	0,1989	0,1585	0,2118	0,1573	0,0718	0,0701	0,0437	0,0349	0,0227	0,0111
	0,138	0,252	0,1911	0,1886	0,1827	0,0742	0,05	0,0271	0,0287	0,0178	0,0104
	0,159	0,2334	0,1476	0,1793	0,1341	0,0507	0,037	0,0164	0,021	0,0103	0,0054
	0,187	0,1277	0,0876	0,1301	0,0723	0,0224	0,035	0,0146	0,0165	0,0081	0,0045
MME	0,05	0,4678	0,4546	0,4479	0,4492	0,434	0,4427	0,4443	0,4424	0,4307	0,4194

Алго-ритм	bpp	CHEN	CC-CHEN	LIU	CC-PEV	CDF	CC-300	CF	JRM	CC-JRM	J+ SRM
	0,1	0,3001	0,2611	0,2574	0,2613	0,2501	0,3026	0,2466	0,2286	0,2091	0,1891
	0,15	0,2165	0,1735	0,1677	0,1721	0,1586	0,2299	0,1608	0,1404	0,1221	0,1027
	0,2	0,0217	0,0104	0,0127	0,0127	0,0124	0,0726	0,0153	0,0112	0,008	0,0059
BCH	0,1	0,4599	0,4496	0,4448	0,4426	0,439	0,4497	0,429	0,4305	0,4229	0,406
	0,2	0,3594	0,3124	0,3087	0,2974	0,2752	0,2958	0,2629	0,2707	0,2369	0,1946
	0,3	0,1383	0,0889	0,0862	0,0779	0,0697	0,0912	0,0663	0,0715	0,0536	0,039
BCHopt	0,1	0,4726	0,4683	0,4558	0,4618	0,4595	0,4684	0,455	0,4515	0,448	0,4306
	0,2	0,4032	0,3712	0,3583	0,3548	0,3368	0,3517	0,3265	0,3253	0,303	0,2582
	0,3	0,24	0,1711	0,1719	0,1605	0,1356	0,1681	0,1289	0,1389	0,1102	0,083
Средняя точность		**73,22%**	**77,05%**	**77,81%**	**77,88%**	**80,77%**	**78,63%**	**82,19%**	**82,98%**	**84,19%**	**85,93 %**

Табл. 2. Сравнение вероятности ошибок стеганоанализа Yedroudj-Net, Xu-Net, Ye-Net и SRM+EC для алгоритмов встраивания WOW и S-UNIWARD при 0,2 bpp и 0,4 bpp
Table 2. Steganalysis error probability comparison of Yedroudj-Net, Xu-Net, Ye-Net, and SRM+EC for embedding algorithms WOW and S-UNIWARD at 0.2 bpp and 0.4 bpp

Модель	BOSS 256x256			
	WOW		S-UNIWARD	
	0.2bpp	0.4bpp	0.2bpp	0.4bpp
SRM+EC	36.5 %	25.5 %	36.6%	24.7 %
Yedroudj	27.8%	14.1%	36.7%	22.8%
Xu-net	32.4 %	20.7 %	39.1 %	27.2 %
Ye-Net	33.1 %	23.2 %	40.0 %	31.2 %

Дополнительное использование RM-модели для обучения статистических классификаторов [2] позволяет увеличить точность классификации. Однако в среднем, при учете отношения количества вложенной информации на пиксель (bpp), значение показателя точности классификации ненамного превосходит точность разработанной модели. Из табл. 2 видно, что разработанная модель не уступает представленным в [3]. В то же время по представленным значениям показателя точности можно сделать вывод о том, что использование нейронных сетей для обнаружения стегоконтейнера, дает лучшие результаты по сравнению с RM-моделью. При этом значение показателя точности разработанной модели может быть увеличено за счет увеличения объема выборки цифровых изображений, используемой на этапе обучения нейронной сети.

6. Заключение

В статье рассмотрена возможность применения аппарата свёрточных нейронных сетей для обнаружения стеганографических вложений в цифровых изображениях. Результаты исследования демонстрируют возможность обнаружения до 85% фактов наличия стеганографических вложений. В ходе проведенного исследования были решены следующие задачи:

- разработана модель свёрточной нейронной сети для обнаружения факта применения стеганографии в цифровых изображениях;
- на основании разработанной модели реализована программа для стегоанализа, реализующая бинарную классификацию цифровых изображений;
- проведен сравнительный анализ использования разработанной модели, статистических классификаторов и других моделей нейронных сетей, используемых для решения

задачи стегоанализа.

К достоинствам предложенного способа обнаружения стеговложений можно отнести достаточную точность и простоту реализации. Реализованная модель позволяет находить скрытые зависимости, не применяя сложных статистических алгоритмов. Из недостатков стоит отметить необходимость решения задачи формирования представительной выборки цифровых изображений, используемой на этапе обучения нейронной сети.

Направлением дальнейших исследований является совершенствование разработанной модели с целью повышения вероятности обнаружения стеганографических вложений и снижения вычислительной сложности ее программной реализации, а также создание собственной базы изображений с различными параметрами вложений.

Список литературы / References

[1]. Boroumand M, Chen M, Fridrich J. Deep Residual Network for Steganalysis of Digital Images. IEEE Transactions on Information Forensics and Security, vol. 14, issue 5, 2019, pp. 1181-1193.

[2]. Kodovsky J, Fridrich J. Rich models for steganalysis of digital images. IEEE Transactions on Information Forensics and Security, vol. 7, issue 3, 2012, pp. 868-882.

[3]. 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, pp. 2092-2096.

[4]. Lerch-Hostalot D, Megias D. Detection of Classiffer Inconsistencies in Image Steganalysis. In Proc. of the ACM Workshop on Information Hiding and Multimedia Security. 2019, pp. 222-229.

[5]. Свёрточная нейронная сеть с нуля. Часть 0. Введение [Электронный ресурс] – Режим доступа: URL: https://programforyou.ru/poleznoe/convolutional-network-from-scratch-part-zero-introduction (01.07.2020) / Convolutional neural network from scratch. Part 0. Introduction. URL: https://programforyou.ru/poleznoe/convolutional-network-from-scratch-part-zero-introduction (in Russian).

[6]. Shafkat I. Intuitively Understanding Convolutions for Deep Learning. URL: https://towardsdatascience.com/intuitively-understanding-convolutions-for-deep-learning-1f6f42faee1 (accessed 06.06.2020).

[7]. Saha S. A Comprehensive Guide to Convolutional Neural Networks – the ELI5 way. URL: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53 (accessed 06.06.2020)

[8]. Yousfi Y, Butora J, Fridrich J, Giboulot Q. Breaking ALASKA: Color Separation for Steganalysis in JPEG Domain. In Proc. of the ACM Workshop on Information Hiding and Multimedia Security, 2019, pp. 138-149.

[9]. Convolution arithmetic tutorial. URL: http://deeplearning.net/software/theano/tutorial/conv_arithmetic.html (accessed 01.07.2020)

[10]. Kodovsky J, Fridrich J. Steganalysis of JPEG images using rich models. In Proc. of the SPIE International Conference on Media Watermarking, Security, and Forensics, 2012, paper 8303-8.

Информация об авторах / Information about authors

Александр Александрович ПОЛУНИН – сотрудник. В его научные интересы входят: машинное обучение, искусственный интеллект на основе нейронных сетей, компьютерное зрение.

Alexander Alexandrovich POLUNIN is an employee. His research interests include machine learning, artificial intelligence based on neural networks, and computer vision.

Элина Андреевна ЯНДАШЕВСКАЯ – сотрудница. К её интересам в научной сфере можно отнести: криптографические методы защиты информации, стеганография, стегоанализ, скрытые каналы передачи данных.

Elina Andreevna YANDASHEVSKAYA is an employee. Her research interests: cryptographic methods of information security, steganography, steganalysis, hidden data transmission channels.