НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
май–июнь 2024 Том 24 № 3 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
May–June 2024 Vol. 24 No 3 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

doi: 10.17586/2226-1494-2024-24-3-490-499

An approach to detecting L_0-optimized attacks on image processing neural networks via means of mathematical statistics

Dmitry A. Esipov✉

ITMO University, Saint Petersburg, 197101, Russian Federation

some1else.d.ma@gmail.com✉, https://orcid.org/0000-0003-4467-5117

Abstract

Artificial intelligence has become widespread in image processing tasks. At the same time, the number of vulnerabilities is increasing in systems implementing these artificial intelligence technologies (the attack surface is increasing). The main threats to information security can be implemented by introducing malicious perturbations into the input data, regardless of their type. To detect such attacks, approaches and methods have been developed based, in particular, on the use of an auto-encoder or the analysis of layers of the target neural network. The disadvantage of existing methods, which significantly reduce the scope of their application, is binding to the dataset or model architecture. This paper discusses the issues of expanding the scope (increasing scalability) of methods for detecting L_0-optimized perturbations introduced by unconventional pixel attacks. An approach to detecting these attacks using statistical analysis of input data, regardless of the model and dataset, is proposed. It is assumed that the pixels of the perturbation embedded in the image, as a result of the L_0-optimized attack, will be considered both local and global outliers. Outlier detection is performed using statistical metrics such as deviation from nearest neighbors and Mahalanobis distance. The evaluation of each pixel (anomaly score) is performed as a product of the specified metrics. A threshold clipping algorithm is used to detect an attack. When a pixel is detected for which the received score exceeds a certain threshold, the image is recognized as distorted. The approach was tested on the CIFAR-10 and MNIST datasets. The developed method has demonstrated high accuracy in detecting attacks. On the CIFAR-10 dataset, the accuracy of detecting onepixel attack (accuracy) was 94.3 %, and when detecting a Jacobian based Saliency Map Attack (JSMA) — 98.3 %. The proposed approach is also applicable in the detection of modified pixels. The proposed approach is applicable for detecting one-pixel attacks and JSMA, but can potentially be used for any L_0-optimized distortions. The approach is applicable for color and grayscale images regardless of the dataset. The proposed approach is potentially universal for the architecture of a neural network, since it uses only input data to detect attacks. The approach can be used to detect images modified by unconventional adversarial attacks in the training sample before the model is formed.

Keywords

artificial neural network, image processing, adversarial attack, pseudonorm L_0, malicious perturbation, one-pixel attack, Jacobian Saliency Map Attack

For citation: Esipov D.A. An approach to detecting L_0-optimized attacks on image processing neural networks via means of mathematical statistics. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2024, vol. 24, no. 3, pp. 490–499. doi: 10.17586/2226-1494-2024-24-3-490-499

УДК 004.056

Подход к обнаружению неконвенциональной пиксельной атаки на нейронные сети обработки изображений методами статистического анализа

Дмитрий Андреевич Есипов✉

Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

some1else.d.ma@gmail.com✉, https://orcid.org/0000-0003-4467-5117

Аннотация

Введение. Искусственный интеллект получил широкое распространение в задачах обработки изображений. Вместе с тем в системах, реализующих технологии искусственного интеллекта, растет количество уязвимостей

(увеличивается поверхность атаки). Основные угрозы информационной безопасности могут быть реализованы посредством внесения вредоносных возмущений во входные данные вне зависимости от их типа. Для обнаружения таких атак были разработаны подходы и методы, основанные, в частности, на применении автокодировщика или анализе слоев целевой нейронной сети. Недостатком существующих методов, значительно снижающих области их применения, является привязка к набору данных или архитектуре модели. В данной работе рассматриваются вопросы расширения областей применения (повышения масштабируемости) методов обнаружения, оптимизированных по псевдонорме L_0 искажений, вносимых неконвенциональными пиксельными атаками. Предложен подход к обнаружению пиксельных атак методами статистического анализа входных данных независимо от модели и набора данных. **Метод.** Предполагается, что пикселы возмущения, встроенные в изображение при адресации атаки, оптимизированной по L_0, будут считаться одновременно и локальными, и глобальными выбросами. Обнаружение выбросов выполняется с использованием таких статистических метрик, как отклонение от ближайших соседей и расстояние Махаланобиса. Оценка каждого пиксела (оценка аномальности) производится как произведение статистических метрик. Для обнаружения атаки применяется алгоритм отсечения по порогу. При обнаружении пиксела, для которого полученная оценка превышает некоторый порог, изображение признается искаженным. **Основные результаты.** Апробация подхода выполнена на наборах данных CIFAR-10 и MNIST. Разработанный метод продемонстрировал высокую точность обнаружения атак. На наборе данных CIFAR-10 точность обнаружения однопиксельной атаки (accuracy) составила 94,3 %, а при обнаружении атаки по карте значимости на основе Якобиана (Jacobian based Saliency Map Attack, JSMA) — 98,3 %. Представленный подход может быть использован в задачах обнаружения искаженных пикселов. **Обсуждение.** Предложенный подход применим для обнаружения однопиксельных атак и JSMA, но потенциально может быть использован для любых искажений, оптимизированных по L_0. Подход применим к цветным изображениям и изображениям в оттенках серого независимо от набора данных. Рассмотренный подход потенциально универсален к архитектуре нейронной сети, поскольку для обнаружения атак использует исключительно входные данные. Подход может быть использован для обнаружения искаженных неконвенциональными пиксельными атаками изображений в обучающей выборке до формирования модели.

Ключевые слова

искусственная нейронная сеть, обработка изображений, состязательная атака, вредоносное возмущение, псевдонорма возмущения L_0, однопиксельная атака, атака по карте значимости на основе Якобиана

Ссылка для цитирования: Есипов Д.А. Подход к обнаружению неконвенциональной пиксельной атаки на нейронные сети обработки изображений методами статистического анализа // Научно-технический вестник информационных технологий, механики и оптики. 2024. Т. 24, № 3. С. 490–499 (на англ. яз.). doi: 10.17586/2226-1494-2024-24-3-490-499

Introduction

Artificial intelligence and machine learning have become widespread due to its significant effectiveness in solving a variety of applied tasks [1]. Neural networks are used for image processing in medical diagnostics [2, 3], biometric authentication [4–6] and in autonomous vehicles [7–9].

At the same time, the use of machine learning and artificial intelligence is associated with characteristic threats. One of these threats is machine learning model evasion[1].

The phenomenon of neural network evasion as a result of an adversarial attack was first demonstrated by Szegedy C. et al. [10] in 2013. Attack methods based on malicious perturbations on neural networks have been continuously improved, methods of disrupting the operation of neural networks in processing various types of data and tasks of the target model have been proposed [1, 11, 12]. Attack algorithms with different characteristics of the introduced perturbation have also been developed.

Attacks based on malicious perturbation. Attacks based on malicious perturbation, including adversarial attacks, involve machine learning model evasion or embedding a backdoor into the specified model by distorting the input data. Evasion involves introducing a perturbation to the input data when using a trained model, in order to embed a backdoor, perturbation of the training dataset is necessary.

The attacks considered are based on the specifics of image processing. Machine learning models, including artificial neural networks, do not see in the understanding familiar to humans. To process images, it performs certain mathematical transformations based on the pixel values of the image. In the learning process, to solve classification tasks, models identify pixel patterns characteristic of a certain class. Elements that have a greater correlation with the target class have greater importance and greater weight. Due to the described specifics of image processing, the introduction of even small perturbations can lead to an incorrect response of the model.

The perturbation introduced by the attack is characterized by distance metrics or norms [1, 11, 12]. Along with such norms as Manhattan distance L_1, Euclid distance L_2 and Chebyshev distance L_∞; the pseudonorm L_0 also used, characterizing the number of elements (pixels) distorted by the attack regardless of the degree of deviation from the original value. It should be noted that the model evasion can be performed by changing only one pixel of the image [13].

The algorithms of generating adversarial examples, characterized by L_0, include one-pixel attack [13], Jacobian Saliency Map Attack (JSMA) [14], Localized and Visible Adversarial Noise (LaVAN) [15], etc.

A one-pixel attack [13] is a neural network evasion attack by perturbating single pixel of the input image. The

[1] MITRE. Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS). Available at: https://atlas.mitre.org/, free access (accessed: 12.12.2023).

attack involves the use of Differential Evolution Algorithm (DEA) [16] to determine the position and value of the embedded pixel. The perturbation introduced by this attack has the lowest value of L_0 since the value of only one-pixel changes. Modifications of the specified attack allow modification of a larger number of pixels.

JSMA [14] is a neural network evasion by perturbating a certain proportion of pixels of the input image. Pixel positions and values are determined in accordance with saliency map created using the neural network forward derivative. This attack involves the distortion of a larger number of pixels than a one-pixel attack; therefore it has a greater impact on the statistical characteristics of the image.

In the current work, L_0-optimized attack algorithms will be considered.

Related works. Due to the relevance of the threat of machine learning model evasion, L_0-optimized attack detection algorithms have been developed [17–21].

OPADA [17] assumes the use of the one-pixel attack itself to protect target model against it. To do this, a set of training data is generated, including both clean images and adversarial examples. The generated set is used to train the classifier based on logistic regression; the responses of the protected neural network are used as predictors. It should be noted that the specified algorithm achieves detection accuracy of 100 % on some neural network architectures, while on others it demonstrates 36.67 %.

Another option to protect against one-pixel attack is to train a variational auto-encoder [18]. The specified defense method involves passing the input data of the target model through a variational auto-encoder trained on the data processed by the model. At the same time, the malicious perturbation introduced by the attack can be eliminated. The considered method achieves 99 % detection and elimination accuracy. However, the auto-encoder allows organizing attack protection only on the dataset on which it was trained.

Wang P. et al. [19] introduced a method for detecting one-pixel attack by analyzing the layers of a neural network and determining the most significant elements of the input data (pixel positions) for each class. The definition of such elements for each class forms a set of coordinates of pixels potentially modified by the attack. Then detection involves checking certain elements in each input image and searching for outliers among the values of these pixels. The presence of an outlier may indicate the fact of an image attack. The accuracy of this method on real data reaches 9.1 %. Since one-pixel attack is addressed in black-box mode, the attacker does not have access to such an investigation of the target of the attack. The position and value of the distorted pixel is determined by the DEA [16]. Then the final coordinates may allow model evasion, but they do not match with those defined when organizing protection. In addition, there may be more than one element for each class that has a significant effect on the model response. Then the attack can be successful if not the most significant element is modified.

Grosse K. et al. [20] introduced an approach to detect attacks by testing statistical hypotheses. Detection is performed by extracting the characteristics of the statistical distribution of image pixels and evaluating these parameters by a trained classifier. This approach makes it possible to detect various types of attacks, including those optimized according to different norms (Fast Gradient Sign Method [22], JSMA [14]). The classifier achieved a JSMA detection accuracy of 83.76 %. One of the limitations of the proposed approach is the strict dependence of the detection quality on the training sample. Therefore, the proposed approach does not allow detecting attacks that are not represented in the mentioned dataset.

Guo et al. [21] used the difference in the response of different models to detect the attack. The proposed approach is based on the possibility of transferring attacks among models due to the similarity of their decision-making boundaries. At the same time, the responses of models trained on the same data may differ in adversarial examples due to differences in the boundaries of decision-making. This phenomenon is called Transferability Prediction Difference. Then the attack marker may be a difference in the response of several models. This detection method allows detecting various attacks. The JSMA detection accuracy of the developed method reaches 97 % on the MNIST[1] dataset and 94 % on CIFAR-10[2]. It should be mentioned that this approach to detection involves significant redundancy, namely the use of several models. Also, when using this method, there is a slight decrease in the quality of the model on undistorted data.

Important disadvantages of existing defense methods are its binding to the architecture of a neural network [16], a dataset [17] or a particular model [18] due to the specifics of the approaches used, which limits their scope of application. Detection by means of mathematical statistics [19] is more universal; however, the proposed approach does not allow detecting various types of L_0-optimized attacks. Then there is a need to develop a more comprehensive approach to detecting the attacks considered, which is the purpose of the current work. The objectives to achieve this purpose are to determine the essence of the proposed approach, develop the algorithms used, design the experiment and evaluate the proposed solution.

Proposed method

The proposed approach involves detection in two stages: image preprocessing and attack detection. Image preprocessing is aimed at calculating the pixel anomaly score. The detection of an attack is assumed by analyzing the data obtained at the previous stage.

The approach is based on the assumption that the pixels of the perturbation will be both local and global outliers. Then their detection consists in the intersection of the sets of specified outliers.

Image preprocessing algorithm. To detect local outliers, the use of deviation from the average of the nearest

[1] Kaggle. MNIST Dataset. Available at: https://www.kaggle.com/datasets/hojjatk/mnist-dataset, free access (accessed: 02.02.2024).

[2] Kaggle. CIFAR-10 — Object Recognition in Images. Available at: https://www.kaggle.com/c/cifar-10/, free access (accessed: 02.02.2024).

neighbors was chosen. It is important to highlight that a distorted pixel can differ significantly in only one of the color components. Then the sum of deviations by color components was chosen as the final deviation estimate:

$$n = \sum_{i=1}^{c} \frac{x_i - \sum_{j=1}^{k} \frac{y_{i,j}}{k}}{255},$$

where n — estimation of deviation from nearest neighbors; i — color channel; c — number of color channels; x_i — value of i-th channel of the pixel being tested; k — number of neighbor pixels; $y_{i,j}$ — value of i-th channel of j-th neighbor pixel.

To determine global outliers, the Mahalanobis distance m [23] from the pixel being tested to the image as a class of pixels was chosen:

$$m = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{S}^{-1} (\mathbf{x} - \boldsymbol{\mu})},$$

where $\mathbf{x}$ — the pixel being tested; $\boldsymbol{\mu}$ – mean pixel value; $\mathbf{S}$ — image pixel covariance matrix.

Z-score (standard score) can also be used to detect global outliers [24]. Since a pixel in a color image is a vector of three elements (RGB component), Z-score is not suitable for it. At the same time, the pixel of a grayscale image has only one value x', so for such images, along with the Mahalanobis distance, a Z-score z can be used:

$$z = \frac{x' - \mu'}{\sigma},$$

where z — Z-score; x' — the pixel being tested in grayscale; μ' — mean pixel value in grayscale; σ — standard pixel deviation.

As an analogy for the intersection of sets of local and global outliers, the product of the estimates obtained can be used. Thus, the final evaluation of a pixel as modified by an attack (anomaly score) is calculated as the product of previously calculated values. Pixel anomaly scores are recorded in a matrix according to their positions in the image. An example of image processing is shown in Fig. 1.

Attack detection algorithm. A cut-off algorithm based on a certain threshold can be used for detection. Then, when a pixel whose anomaly score exceeds the specified value is detected, the algorithm determines the image as attacked. The value of the cut-off threshold is the only parameter of the algorithm. It should be noted that for this algorithm it is not necessary to have a complete matrix of pixel anomaly scores. A comparison with the threshold for each pixel can be performed immediately after calculating its anomaly score. In the case of forming a complete matrix, cutting off the threshold will show the positions of the pixels distorted by the attack.

Further in the text, a combination of the above algorithms will be used as the L_0-optimized attack detection method. It should be noted that other attack detection algorithm options can be used.

Design of the experiment

Attack algorithms. Two attack algorithms were chosen as L_0-optimized attacks: one-pixel attack and JSMA. It should be emphasized that the proposed approach is potentially applicable to other L_0-optimized attacks.

An open access program code was used to address the one-pixel attack[1]. JSMA was performed using the advertorch[2] library of the Python programming language.

Datasets used. According to [13], one-pixel attack is effective for low-resolution images (up to 65 × 65 pixels). Therefore, three sets of images satisfying the specified limitation were used to conduct the experiment.

CIFAR-10[3] contains 60,000 color images with a resolution of 32 × 32 pixels, pertaining to 10 classes. The specified dataset was used to evaluate the performance of the proposed approach on color images.

MNIST[4] contains 60,000 grayscale images with a resolution of 28 × 28 pixels, corresponding to numbers from 0 to 9, that is, 10 classes. The specified dataset was used to evaluate the approach performance on grayscale images. It is important to note that the images in MNIST have significant contrast, which is why they are close to black and white images. At that time, the discolored

[1] Github. DebangLi / one-pixel-attack-pytorch. Available at: https://github.com/DebangLi/one-pixel-attack-pytorch/tree/master, free access (accessed: 30.09.2023).

[2] Github. BorealisAI / advertorch. Available at: https://github.com/BorealisAI/advertorch, free access (accessed: 30.09.2023).

[3] Kaggle. CIFAR-10 — Object Recognition in Images. Available at: https://www.kaggle.com/c/cifar-10/, free access (accessed: 02.02.2024).

[4] Kaggle. MNIST Dataset. Available at: https://www.kaggle.com/datasets/hojjatk/mnist-dataset, free access (accessed: 02.02.2024).

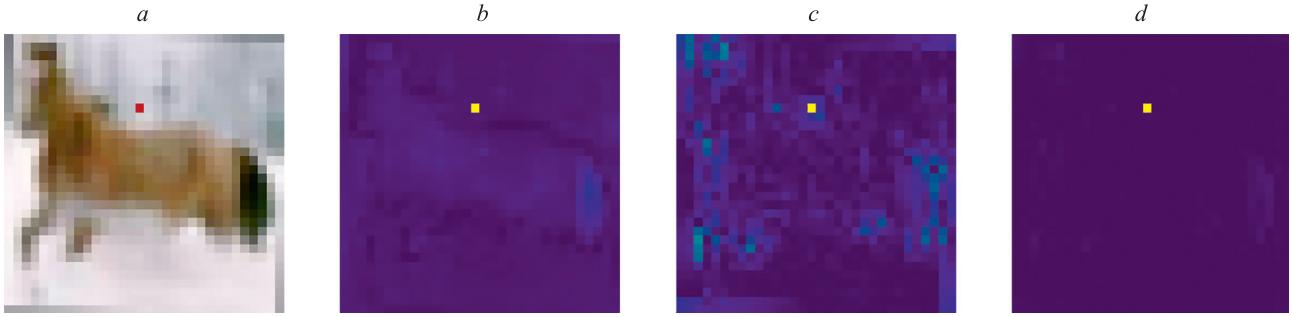

Fig. 1. An example of image processing: initial image (*a*); matrix of Mahalanobis distances (*b*); matrix of deviations from nearest neighbors (*c*); matrix of pixel final evaluation (*d*)

Table 1. Characteristics of the datasets used

Dataset	Attack algorithm		Number of elements	
	name	gamma, %	total	used for evaluation
CIFAR-10	One-pixel attack	—	10,945	10,000
	JSMA	1	17,098	10,000
		3	38,991	10,000
		5	48,709	10,000
CIFAR-10-G	One-pixel attack	—	17,353	10,000
	JSMA	1	35,126	10,000
		3	39,778	10,000
		5	44,106	10,000
MNIST	One-pixel attack	—	7472	2000
	JSMA	1	2081	2000
		3	13,390	2000
		5	24,404	2000

CIFAR-10 (CIFAR-10-G) was also used to evaluate the approach on grayscale images.

The characteristics of the obtained sets of adversarial examples are shown in Table 1.

The sets of perturbated images used in further experiments, as well as the attacked neural networks, are available on GitHub[1].

Evaluation metrics. Accuracy shows the proportion of correct responses of the algorithm, regardless of the type of error. Precision shows the ability of the algorithm to distinguish objects of a certain class from objects of other classes, thereby taking into account only type-I errors. Recall determines the possibility of identifying objects of a certain class by an algorithm and takes into account type-II errors. When detecting the fact of an attack, it is necessary to consider both types of errors separately; therefore, the F1-score was chosen to evaluate the algorithm. Accuracy was also calculated for comparison with analogues.

Since the perturbation introduced by the attacks takes up a small fraction of the pixels of the image, there will be significant disparity of classes, which does not allow the use of the accuracy metric. Then the metrics precision, recall and F1-score can be calculated. The F1-score was chosen for the final evaluation of the perturbation detection.

Then F1-score and accuracy will be used as quality indicators to detect an attack and F1-score to detect a perturbation.

Results and analysis

Determining the cut-off threshold for detecting an attack on color images. To determine the value of the cut-off threshold corresponding to the highest indicator of the F1-score of attack detection, the proposed method with different values of the cut-off threshold was applied to the formed dataset. Since the values of the Mahalanobis distance and the deviation from the nearest neighbors are non-negative, their product is also non-negative. Then the values of the cut-off threshold were selected from the range from 0 to 10 in increments of 0.01 (Fig. 2). The value of the cut-off threshold at which the highest values of the evaluation metrics are achieved is shown in Table 2.

According to Table 2, the L_0-optimized attack detection method based on the proposed approach demonstrates high quality indicators of one-pixel attack detection and JSMA. The approach can also be used to detect other similar attacks. It should be noted that for various attacks, the maximum value of the evaluation metrics is observed at different values of the cut-off threshold, which does not allow detecting various attacks simultaneously. An option to eliminate this shortcoming is to use a different approach to detection. Then the statistical distribution of the obtained pixel anomaly scores and the characteristics of this distribution can be used.

Determining the algorithms used and the cut-off threshold for detecting an attack on grayscale images. Determining the threshold for detecting an L_0-optimized attack on grayscale images was performed similarly to the previous step. In addition to the values of the cut-off thresholds, the application of various algorithms for detecting global outliers, namely calculating the Mahalanobis distance and Z-score, was also considered. The obtained results of detecting attacks on the CIFAR-10-G dataset are shown in Table 3, on MNIST — in Table 4.

According to Table 3, the choice of an algorithm for detecting global outliers does not significantly affect the quality of attack detection. The method also demonstrates high detection quality indicators on the CIFAR-10-G dataset.

According to Table 4, there is a significant decline in quality indicators in high-contrast images, due to limitations of global outlier detection algorithms. Then the developed method and the proposed approach have a limited scope of application in images with high contrast.

[1] GitHub. iNDm3802 / L0-optimized_attack_detection. Available at: https://github.com/iNDm3802/L0-optimized_attack_detection, free access (accessed: 01.03.2023).

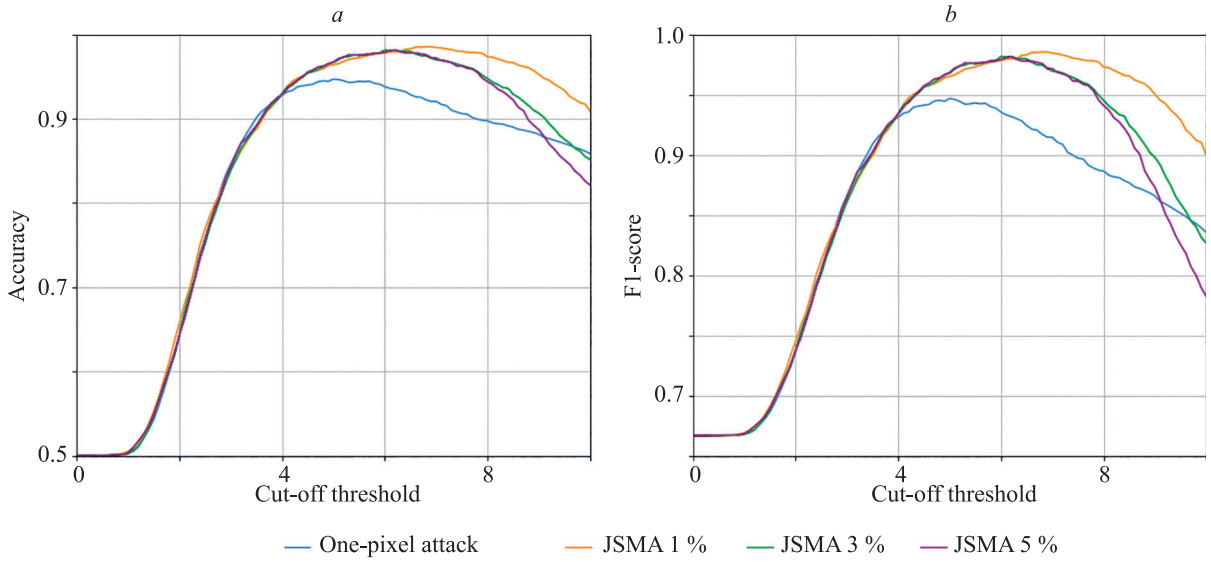

Fig. 2. Dependence of quality indicators on the cut-off threshold: accuracy (*a*); F1-score (*b*)

Table 2. The highest quality indicators for detecting attacks on color images (CIFAR-10)

Attack		Cut-off threshold	Accuracy, %	F1-score, %
One-pixel attack		5.05	94.27	94.27
JSMA	gamma = 1 %	6.55	98.32	98.32
	gamma = 3 %	6.27	98.06	98.07
	gamma = 5 %	6.08	98.11	98.12

Table 3. The highest quality indicators for detecting attacks on grayscale images (CIFAR-10-G)

Attack		Algorithm	Cut-off threshold	Accuracy, %	F1-score, %
One-pixel attack		Mahalanobis distance	1.07	88.67	88.76
		Z-score	1.07	88.68	88.77
JSMA	gamma = 1 %	Mahalanobis distance	1.44	94.10	93.96
		Z-score	1.43	94.09	93.95
	gamma = 3 %	Mahalanobis distance	1.37	94.20	94.12
		Z-score	1.38	94.21	94.12
	gamma = 5 %	Mahalanobis distance	1.35	94.10	94.04
		Z-score	1.35	94.10	94.05

Table 4. The highest quality indicators for detecting attacks on grayscale images (MNIST)

Attack		Algorithm	Cut-off threshold	Accuracy, %	F1-score, %
One-pixel attack		Mahalanobis distance	2.00	80.05	80.53
		Z-score	1.99	79.93	80.50
JSMA	gamma = 1 %	Mahalanobis distance	1.40	62.25	69.81
		Z-score	1.40	62.20	69.78
	gamma = 3 %	Mahalanobis distance	1.51	67.15	72.41
		Z-score	1.51	67.15	72.42
	gamma = 5 %	Mahalanobis distance	1.35	62.95	71.42
		Z-score	1.37	63.25	71.45

Determination of the cut-off threshold for perturbation detection. The determination of the threshold for detecting the perturbation introduced by the considered attacks was performed similarly to the previous sections (Fig. 3, Table 5). It should be noted that the perturbation detection was performed only on the attacked images, which is possible only after determining the fact of the attack. Both the initial value of the pixel anomaly scores and the normalized value were also used. The CIFAR-10 dataset was used to detect the perturbation.

According to Table 5, the method also demonstrates high quality indicators for perturbation detection in color images. It should be noted that in order to detect a perturbation characteristic of a one-pixel attack, a higher value of the F1-score is achieved using normalized values of the anomaly score. At the same time, to detect the perturbation characteristic of JSMA, a greater value of the F1-score is observed when using the initial values. It should be noted that the difference in JSMA perturbation detection quality when using both types of values is limited. Then, in order to detect a perturbation, an attack classification should be performed, which is also possible by analyzing the statistical distribution of the obtained pixel anomaly scores and its characteristics.

Data on the detection of perturbation on other datasets, as well as more complete information about the results of the experiment, are available on GitHub[1].

Performance evaluation. The preprocessing algorithm, like the detection algorithm, has a linear computational complexity of O(n), where n corresponds to the number of pixels of the image taking into account the number of color channels, that is, its shape.

Performance evaluation of the L_0-optimized attack detection method based on the proposed approach is shown in Table 6. Calculations were performed on the following hardware:

— CPU: Intel(R) Core(TM) i7-10700 CPU @ 2.90GHz, 2904 MHz, cores: 8, logical processors: 16;
— RAM: 32.0 GB.

According to Table 6, the method demonstrated processing speeds from 17.7 to 46.7 images per second for CIFAR-10 and MNIST, respectively, depending on their characteristics.

[1] GitHub. iNDm3802 / L0-optimized_attack_detection. Available at: https://github.com/iNDm3802/L0-optimized_attack_detection, free access (accessed: 01.03.2023).

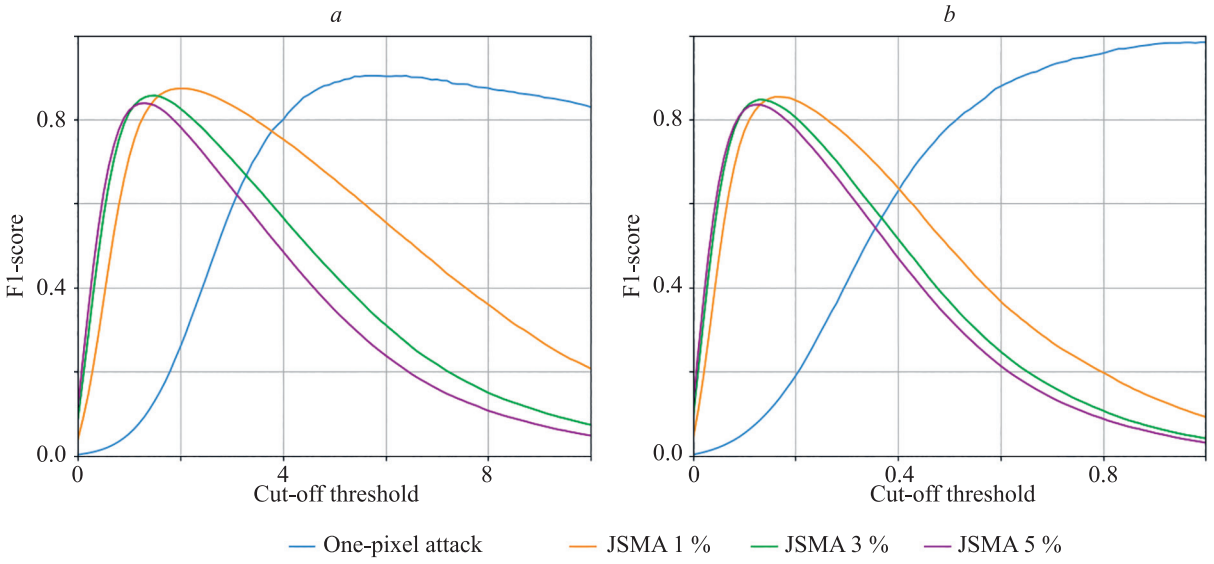

Fig. 3. Dependence of F1-score vs. the cut-off threshold when using: initial values (*a*); normalized values (*b*)

Table 5. The highest quality indicators for detecting perturbation on color images (CIFAR-10)

Attack		Anomaly score value	Cut-off threshold	F1-score, %
One-pixel attack		Initial	6.06	91.10
		Normalized	1.00	98.24
JSMA	gamma = 1 %	Initial	2.00	88.23
		Normalized	0.17	85.58
	gamma = 3 %	Initial	1.46	85.43
		Normalized	0.13	84.41
	gamma = 5 %	Initial	1.30	83.64
		Normalized	0.13	83.20

Table 6. Performance evaluation of the method

Dataset	Count of images	Image shape	Time, s	
			Total	Per image
CIFAR-10	10,000	3 color channels, 32 × 32 pixels	564.966	0.056
CIFAR-10-G		1 color channel, 32 × 32 pixels	272.771	0.027
MNIST		1 color channel, 28 × 28 pixels	214.132	0.021

Table 7. Comparative analysis of L_0-optimized attack detection methods

Method	Dataset	Attack		Accuracy, %
OPADA [17]	CIFAR-10	One-pixel attack		36.67–100
Alatalo J. et al. [18]	TUPAC16	One-pixel attack		99
Wang P. et al. [19]	CIFAR-10	One-pixel attack		9.1
Grosse K. et al. [20]	MNIST	JSMA		83.76
Guo F. et al. [21]	CIFAR-10	JSMA		94
	MNIST			97
Developed	CIFAR-10	One-pixel attack		94.27
		JSMA	gamma = 1 %	98.32
			gamma = 3 %	98.06
			gamma = 5 %	98.11
	MNIST	One-pixel attack		80.05
		JSMA	gamma = 1 %	62.25
			gamma = 3 %	67.15
			gamma = 5 %	62.95

Discussion

A comparative analysis of the developed L_0-optimized attack detection method based on the proposed approach with analogues is shown in Table 7.

According to the comparative analysis, the developed method demonstrates quality indicators comparable to analogues. However, unlike analogues, the method and the proposed approach is not bound to either a dataset or the architecture of a neural network, or to the presence of a trained model. Then it can be used to detect perturbated images in the training sample [25]. In addition, it allows detecting both the fact of an attack and the pixels modified by the attack. The method and the approach are also applicable to various L_0-optimized attacks.

The developed method has the following limitations: different values of the cut-off threshold for different attacks, the need to classify attacks in order to detect perturbation. To eliminate these limitations, it is necessary to develop another attack detection algorithm based on the obtained pixel anomaly scores. Also, due to the use of the Mahalanobis distance, the disadvantage of the developed method and the proposed approach is a decline in attack detection quality on contrasting images.

Conclusion

The proposed approach allows detecting the fact of an attack based on L_0-optimized perturbation, as well as the perturbation introduced by the specified attack. The method based on the approach demonstrates high quality indicators when detecting one-pixel attack and JSMA and can be used to detect other similar attacks. The approach is bound neither to a dataset, nor to the architecture of a neural network, nor to the presence of a trained model, which is why it can be used to detect distorted images in a training sample.

The direction of further work is to develop an algorithm for detecting attacks based on the obtained pixel anomaly scores, namely by analyzing the statistical distribution of the obtained values and its characteristics. Another direction is to verify the applicability and possible modification of the developed method for detecting an attack by embedding an adversarial patch.

References

1. Esipov D.A., Buchaev A.Y., Kerimbay A., Puzikova Y.V., Saidumarov S.K., Sulimenko N.S., Popov I.Yu., Karmanovskiy N.S. Attacks based on malicious perturbations on image processing systems and defense methods against them. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2023, vol. 23, no. 4, pp. 720–733. (in Russian). https://doi.org/10.17586/2226-1494-2023-23-4-720-733
2. Sarvamangala D.R., Kulkarni R.V. Convolutional neural networks in medical image understanding: a survey. *Evolutionary Intelligence*, 2022, vol. 15, no. 1, pp. 1–22. https://doi.org/10.1007/s12065-020-00540-3
3. Mahmood M., Al-Khateeb B., Alwash W. A review on neural networks approach on classifying cancers. *IAES International Journal of Artificial Intelligence*, 2020, vol. 9, no. 2, pp. 317–326. https://doi.org/10.11591/ijai.v9.i2.pp317-326
4. Almabdy S., Elrefaei L. Deep convolutional neural network-based approaches for face recognition. *Applied Sciences*, 2019, vol. 9, no. 20, pp. 4397. https://doi.org/10.3390/app9204397
5. Khan M.Z., Harous S., Hassan S.U., Khan M.U.G., Iqbal R., Mumtaz S. Deep unified model for face recognition based on convolution neural network and edge computing. *IEEE Access*, 2019, vol. 7, pp. 72622–72633. https://doi.org/10.1109/ACCESS.2019.2918275
6. Zhang Y., Shi D., Zhan X., Cao D., Zhu K., Li Z. Slim-ResCNN: A deep residual convolutional neural network for fingerprint liveness detection. *IEEE Access*, 2019, vol. 7, pp. 91476–91487. https://doi.org/10.1109/ACCESS.2019.2927357
7. Severino A., Curto S., Barberi S., Arena F., Pau G. Autonomous vehicles: an analysis both on their distinctiveness and the potential impact on urban transport systems. *Applied Sciences*, 2021, vol. 11, no. 8, pp. 3604. https://doi.org/10.3390/app11083604
8. Wang L., Fan X., Chen J., Cheng J., Tan J., Ma X. 3D object detection based on sparse convolution neural network and feature fusion for autonomous driving in smart cities. *Sustainable Cities and Society*, 2020, vol. 54, pp. 102002. https://doi.org/10.1016/j.scs.2019.102002
9. Chen L., Lin S., Lu X., Cao D., Wu H., Guo C., Liu C., Wang F.Y. Deep neural network based vehicle and pedestrian detection for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 2021, vol. 22. no. 6, pp. 3234–3246. https://doi.org/10.1109/TITS.2020.2993926
10. Szegedy C., Zaremba W., Sutskever I., Bruna J., Erhan D., Goodfellow I., Fergus R. Intriguing properties of neural networks. *arXiv*, 2013, arXiv:1312.6199, https://doi.org/10.48550/arXiv.1312.6199
11. Akhtar N., Mian A., Kardan N., Shah M. Advances in adversarial attacks and defenses in computer vision: A survey. *IEEE Access*, 2021, vol. 9, pp. 155161–155196. https://doi.org/10.1109/ACCESS.2021.3127960
12. Huang X., Kroening D., Ruan W., Sharp J., Sun Y., Thamo E., Wu M., Yi X. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer Science Review*, 2020, vol. 37, pp. 100270. https://doi.org/10.1016/j.cosrev.2020.100270
13. Su J., Vargas D.V., Sakurai K. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 2019, vol. 23, no. 5, pp. 828–841. https://doi.org/10.1109/TEVC.2019.2890858
14. Papernot N., McDaniel P., Jha S., Fredrikson M., Celik Z.B., Swami A. The limitations of deep learning in adversarial settings. *Proc. of the 2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, 2016, pp. 372–387. https://doi.org/10.1109/EuroSP.2016.36
15. Karmon D., Zoran D., Goldberg Y. Lavan: Localized and visible adversarial noise. *Proceedings of Machine Learning Research*, 2018, vol. 80, pp. 2507–2515.
16. Das S., Suganthan P.N. Differential evolution: A survey of the state-of-the-art. *IEEE Transactions on Evolutionary Computation*, 2011, vol. 15, no. 1, pp. 4–31. https://doi.org/10.1109/TEVC.2010.2059031
17. Nguyen-Son H.Q., Thao T.P., Hidano S., Bracamonte V., Kiyomoto S., Yamaguchi R.S. OPA2D: One-pixel attack, detection, and defense in deep neural networks. *Proc. of the 2021 International Joint Conference on Neural Networks (IJCNN)*, 2021, pp. 1–10. https://doi.org/10.1109/IJCNN52387.2021.9534332

Литература

1. Есипов Д.А., Бучаев А.Я., Керимбай А., Пузикова Я.В., Сайдумаров С.К., Сулименко Н.С., Попов И.Ю., Кармановский Н.С. Атаки на основе вредоносных возмущений на системы обработки изображений и методы защиты от них // Научно-технический вестник информационных технологий, механики и оптики. 2023. Т. 23. № 4. С. 720–733. https://doi.org/10.17586/2226-1494-2023-23-4-720-733
2. Sarvamangala D.R., Kulkarni R.V. Convolutional neural networks in medical image understanding: a survey // Evolutionary Intelligence. 2022. V. 15. N 1. P. 1–22. https://doi.org/10.1007/s12065-020-00540-3
3. Mahmood M., Al-Khateeb B., Alwash W. A review on neural networks approach on classifying cancers // IAES International Journal of Artificial Intelligence. 2020. V. 9. N 2. P. 317–326. https://doi.org/10.11591/ijai.v9.i2.pp317-326
4. Almabdy S., Elrefaei L. Deep convolutional neural network-based approaches for face recognition // Applied Sciences. 2019. V. 9. N 20. P. 4397. https://doi.org/10.3390/app9204397
5. Khan M.Z., Harous S., Hassan S.U., Khan M.U.G., Iqbal R., Mumtaz S. Deep unified model for face recognition based on convolution neural network and edge computing // IEEE Access. 2019. V. 7. P. 72622–72633. https://doi.org/10.1109/ACCESS.2019.2918275
6. Zhang Y., Shi D., Zhan X., Cao D., Zhu K., Li Z. Slim-ResCNN: A deep residual convolutional neural network for fingerprint liveness detection // IEEE Access. 2019. V. 7. P. 91476–91487. https://doi.org/10.1109/ACCESS.2019.2927357
7. Severino A., Curto S., Barberi S., Arena F., Pau G. Autonomous vehicles: an analysis both on their distinctiveness and the potential impact on urban transport systems // Applied Sciences. 2021. V. 11. N 8. P. 3604. https://doi.org/10.3390/app11083604
8. Wang L., Fan X., Chen J., Cheng J., Tan J., Ma X. 3D object detection based on sparse convolution neural network and feature fusion for autonomous driving in smart cities // Sustainable Cities and Society. 2020. V. 54. P. 102002. https://doi.org/10.1016/j.scs.2019.102002
9. Chen L., Lin S., Lu X., Cao D., Wu H., Guo C., Liu C., Wang F.Y. Deep neural network based vehicle and pedestrian detection for autonomous driving: A survey // IEEE Transactions on Intelligent Transportation Systems. 2021. V. 22. N 6. P. 3234–3246. https://doi.org/10.1109/TITS.2020.2993926
10. Szegedy C., Zaremba W., Sutskever I., Bruna J., Erhan D., Goodfellow I., Fergus R. Intriguing properties of neural networks // arXiv. 2013. arXiv:1312.6199. https://doi.org/10.48550/arXiv.1312.6199
11. Akhtar N., Mian A., Kardan N., Shah M. Advances in adversarial attacks and defenses in computer vision: A survey // IEEE Access. 2021. V. 9. P. 155161–155196. https://doi.org/10.1109/ACCESS.2021.3127960
12. Huang X., Kroening D., Ruan W., Sharp J., Sun Y., Thamo E., Wu M., Yi X. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability // Computer Science Review. 2020. V. 37. P. 100270. https://doi.org/10.1016/j.cosrev.2020.100270
13. Su J., Vargas D.V., Sakurai K. One pixel attack for fooling deep neural networks // IEEE Transactions on Evolutionary Computation. 2019. V. 23. N 5. P. 828–841. https://doi.org/10.1109/TEVC.2019.2890858
14. Papernot N., McDaniel P., Jha S., Fredrikson M., Celik Z.B., Swami A. The limitations of deep learning in adversarial settings // Proc. of the 2016 IEEE European Symposium on Security and Privacy (EuroS&P). 2016. P. 372–387. https://doi.org/10.1109/EuroSP.2016.36
15. Karmon D., Zoran D., Goldberg Y. Lavan: Localized and visible adversarial noise // Proceedings of Machine Learning Research. 2018. V. 80. P. 2507–2515.
16. Das S., Suganthan P.N. Differential evolution: A survey of the state-of-the-art // IEEE Transactions on Evolutionary Computation. 2011. V. 15. N 1. P. 4–31. https://doi.org/10.1109/TEVC.2010.2059031
17. Nguyen-Son H.Q., Thao T.P., Hidano S., Bracamonte V., Kiyomoto S., Yamaguchi R.S. OPA2D: One-pixel attack, detection, and defense in deep neural networks // Proc. of the 2021 International Joint Conference on Neural Networks (IJCNN). 2021. P. 1–10. https://doi.org/10.1109/IJCNN52387.2021.9534332
18. Alatalo J., Sipola T., Kokkonen T. Detecting one-pixel attacks using variational autoencoders // Lecture Notes in Networks and Systems.

18. Alatalo J., Sipola T., Kokkonen T. Detecting one-pixel attacks using variational autoencoders. *Lecture Notes in Networks and Systems*, 2022, vol. 468, pp. 611–623. https://doi.org/10.1007/978-3-031-04826-5_60
19. Wang P., Cai Z., Kim D., Li W. Detection mechanisms of one-pixel attack. *Wireless Communications and Mobile Computing*, 2021, vol. 2021, pp. 1–8. https://doi.org/10.1155/2021/8891204
20. Grosse K., Manoharan P., Papernot N., Backes M., McDaniel P. On the (statistical) detection of adversarial examples. *arXiv*, 2017, arXiv:1702.06280, https://doi.org/10.48550/arXiv.1702.06280
21. Guo F., Zhao Q., Li X., Kuang X., Zhang J., Han Y., Tan Y.A. Detecting adversarial examples via prediction difference for deep neural networks. *Information Sciences*, 2019, vol. 501, pp. 182–192. https://doi.org/10.1016/j.ins.2019.05.084
22. Goodfellow I.J., Shlens J., Szegedy C. Explaining and harnessing adversarial examples. *arXiv*, 2014, arXiv:1412.6572. https://doi.org/10.48550/arXiv.1412.6572
23. McLachlan G.J. Mahalanobis distance. *Resonance*, 1999, vol. 4, no. 6, pp. 20–26. https://doi.org/10.1007/bf02834632
24. Curtis A.E., Smith T.A., Ziganshin B.A., Elefteriades J.A. The mystery of the Z-score. *Aorta*, 2016, vol. 4, no. 4, pp. 124–130. https://doi.org/10.12945/j.aorta.2016.16.014
25. Zhong H., Liao C., Squicciarini A., Zhu S., Miller D. Backdoor embedding in convolutional neural network models via invisible perturbation. *Proc. of the Tenth ACM Conference on Data and Application Security and Privacy*, 2020, pp. 97–108. https://doi.org/10.1145/3374664.3375751

2022. V. 468. P. 611–623. https://doi.org/10.1007/978-3-031-04826-5_60
19. Wang P., Cai Z., Kim D., Li W. Detection mechanisms of one-pixel attack // Wireless Communications and Mobile Computing. 2021. V. 2021. P. 1–8. https://doi.org/10.1155/2021/8891204
20. Grosse K., Manoharan P., Papernot N., Backes M., McDaniel P. On the (statistical) detection of adversarial examples // arXiv. 2017. arXiv:1702.06280. https://doi.org/10.48550/arXiv.1702.06280
21. Guo F., Zhao Q., Li X., Kuang X., Zhang J., Han Y., Tan Y.A. Detecting adversarial examples via prediction difference for deep neural networks // Information Sciences. 2019. V. 501. P. 182–192. https://doi.org/10.1016/j.ins.2019.05.084
22. Goodfellow I.J., Shlens J., Szegedy C. Explaining and harnessing adversarial examples // arXiv. 2014. arXiv:1412.6572. https://doi.org/10.48550/arXiv.1412.6572
23. McLachlan G.J. Mahalanobis distance // Resonance. 1999. V. 4. N 6. P. 20–26. https://doi.org/10.1007/bf02834632
24. Curtis A.E., Smith T.A., Ziganshin B.A., Elefteriades J.A. The mystery of the Z-score // Aorta. 2016. V. 4. N 4. P. 124–130. https://doi.org/10.12945/j.aorta.2016.16.014
25. Zhong H., Liao C., Squicciarini A., Zhu S., Miller D. Backdoor embedding in convolutional neural network models via invisible perturbation // Proc. of the Tenth ACM Conference on Data and Application Security and Privacy. 2020. P. 97–108. https://doi.org/10.1145/3374664.3375751

Author

Dmitry A. Esipov — Assistant, ITMO University, Saint Petersburg, 197101, Russian Federation, sc 57954958600, https://orcid.org/0000-0003-4467-5117, some1else.d.ma@gmail.com

Автор

Есипов Дмитрий Андреевич — ассистент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, sc 57954958600, https://orcid.org/0000-0003-4467-5117, some1else.d.ma@gmail.com

Received 11.03.2024
Approved after reviewing 25.04.2024
Accepted 24.05.2024

Статья поступила в редакцию 11.03.2024
Одобрена после рецензирования 25.04.2024
Принята к печати 24.05.2024