НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
март–апрель 2024 Том 24 № 2 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
March–April 2024 Vol. 24 No 2 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

doi: 10.17586/2226-1494-2024-24-2-256-266

A new method for countering evasion adversarial attacks on information systems based on artificial intelligence

Alisa A. Vorobeva[1]✉, Maxim A. Matuzko[2], Dmitry I. Sivkov[3], Roman I. Safiullin[4], Alexander A. Menshchikov[5]

[1,2,3,4,5] ITMO University, Saint Petersburg, 197101, Russian Federation

[1] vorobeva@itmo.ru✉, https://orcid.org/0000-0001-6691-6167
[2] mxmmtzk@gmail.com, https://orcid.org/0009-0006-2179-6847
[3] 485280@mail.ru, https://orcid.org/0000-0002-3008-3789
[4] romsaaf@mail.ru, https://orcid.org/0009-0004-8635-9432
[5] menshikov@itmo.ru, https://orcid.org/0000-0002-2287-4310

Abstract

Modern artificial intelligence (AI) technologies are being used in a variety of fields, from science to everyday life. However, the widespread use of AI-based systems has highlighted a problem with their vulnerability to adversarial attacks. These attacks include methods of fooling or misleading an artificial neural network, disrupting its operations, and causing it to make incorrect predictions. This study focuses on protecting image recognition models against adversarial evasion attacks which have been recognized as the most challenging and dangerous. In these attacks, adversaries create adversarial data that contains minor perturbations compared to the original image, and then send it to a trained model in an attempt to change its response to the desired outcome. These distortions can involve adding noise or even changing a few pixels. In this paper, we consider the most relevant methods for generating adversarial data: the Fast Gradient Sign Method (FGSM), the Square Method (SQ), the predicted gradient descent method (PGD), the Basic Iterative Method (BIM), the Carlini-Wagner method (CW) and Jacobian Saliency Map Attack (JSMA). We also study modern techniques for defending against evasion attacks through model modification, such as adversarial training and pre-processing of incoming data, including spatial smoothing, feature squeezing, jpeg compression, minimizing total variance, and defensive distillation. While these methods are effective against certain types of attacks, to date, there is no single method that can be used as a universal defense. Instead, we propose a new method that combines adversarial learning with image pre-processing. We suggest that adversarial training should be performed on adversarial samples generated from common attack methods which can then be effectively defended against. The image preprocessing aims to counter attacks that were not considered during adversarial training. This allows to protect the system from new types of attacks. It is proposed to use jpeg compression and feature squeezing on the pre-processing stage. This reduces the impact of adversarial perturbations and effectively counteracts all types of considered attacks. The evaluation of image recognition model (based on convolutional neural network) performance metrics based was conducted. The experimental data included original images and adversarial images created using attack FGSM, PGD, BIM, SQ, CW, and JSMA methods. At the same time, adversarial training of the model was performed in experiments on data containing only adversarial examples for the FGSM, PGD, and BIM attack methods. Dataset used in experiments was balanced. The average accuracy of image recognition was estimated with crafted adversarial imaged datasets. It was concluded that adversarial training is effective only in countering attacks that were used during model training, while methods of pre-processing incoming data are effective only against more simple attacks. The average recognition accuracy using the developed method was 0.94, significantly higher than those considered methods for countering attacks. It has been shown that the accuracy without using any counteraction methods is approximately 0.19, while with adversarial learning it is 0.79. Spatial smoothing provides an accuracy of 0.58, and feature squeezing results in an accuracy of 0.88. Jpeg compression provides an accuracy of 0.37, total variance minimization — 0.58 and defensive distillation — 0.44. At the same time, image recognition accuracy provided by developed method for FGSM, PGD, BIM, SQ, CW, and JSMA attacks is 0.99, 0.99, 0.98, 0.98, 0.99 and 0.73, respectively. The developed method is a more universal solution for countering all types of attacks and works quite effectively against complex adversarial attacks such as CW and JSMA. The developed method makes it possible to increase accuracy of image recognition model for adversarial images. Unlike adversarial learning, it also increases recognition accuracy on adversarial data generated using attacks not used on training stage. The results are useful for researchers and practitioners in the field of machine learning.

Keywords
machine learning methods, adversarial attacks, defense mechanisms, AI-based information systems, adversarial learning

For citation: Vorobeva A.A., Matuzko M.A., Sivkov D.I., Safiullin R.I., Menshchikov A.A. A new method for countering evasion adversarial attacks on information systems based on artificial intelligence. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2024, vol. 24, no. 2, pp. 256–266. doi: 10.17586/2226-1494-2024-24-2-256-266

УДК 004.056

Новый метод противодействия состязательным атакам уклонения на информационные системы, основанные на искусственном интеллекте

Алиса Андреевна Воробьева[1]✉, Максим Александрович Матузко[2], Дмитрий Игоревич Сивков[3], Роман Ильшатович Сафиуллин[4], Александр Алексеевич Менщиков[5]

[1,2,3,4,5] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

[1] vorobeva@itmo.ru✉, https://orcid.org/0000-0001-6691-6167
[2] mxmmtzk@gmail.com, https://orcid.org/0009-0006-2179-6847
[3] 485280@mail.ru, https://orcid.org/0000-0002-3008-3789
[4] romsaaf@mail.ru, https://orcid.org/0009-0004-8635-9432
[5] menshikov@itmo.ru, https://orcid.org/0000-0002-2287-4310

Аннотация

Введение. Современные технологии искусственного интеллекта находят применение в различных областях науки и повседневной жизни. Повсеместное внедрение систем, основанных на методах искусственного интеллекта, выявило проблему их уязвимости перед состязательными атаками, включающими методы обмана искусственной нейронной сети и нарушения ее работы. В работе основное внимание уделено защите моделей распознавания изображений от состязательных атак уклонения, признанных в настоящее время наиболее опасными. При таких атаках создаются состязательные данные, содержащие незначительные искажения относительно исходных, и происходит отправка их на обученную модель с целью изменения ее «ответа» на вариант, необходимый злоумышленнику. Искажения могут включать добавление шума или изменение нескольких пикселов изображения. Рассмотрены наиболее актуальные подходы к созданию состязательных данных: метод быстрого градиента (Fast Gradient Sign Method, FGSM), метод квадрата (Square Method, SQ), метод прогнозируемого градиентного спуска (Predicted Gradient Descent, PGD), базовый итеративный метод (Basic Iterative Method, BIM), метод Карлини и Вагнера (Carlini-Wagner, CW), метод карт значимости Якобиана (Jacobian Saliency Map Attack, JSMA). Исследованы современные методы противодействия атакам уклонения, основанные на модификации модели — состязательное обучение и предварительная обработка поступающих данных: пространственное сглаживание, сжатие признаков, JPEG-сжатие, минимизация общей дисперсии, оборонительная дистилляция. Эти методы эффективны только против определенных видов атак. На сегодняшний день ни один метод противодействия не может быть применен в качестве универсального решения. **Метод.** Предложен новый метод, сочетающий состязательное обучение с предварительной обработкой изображений. Состязательное обучение выполнено на основе состязательных данных, создаваемых с распространенных атак, что позволяет эффективно им противодействовать. Предварительная обработка изображений предназначена для противодействия атакам, которые не учитывались при состязательном обучении, что дает возможность защитить систему от атак новых типов. Обработка осуществлена методом JPEG-сжатия и сжатия признаков для уменьшения влияния состязательных искажений и более эффективного противодействия всем видам рассмотренных атак. **Основные результаты.** Проведена оценка показателей качества распознавания изображений на основе искусственной нейронной сети. Экспериментальные данные включали оригинальные и измененные изображения, созданные с использованием методов атак типов FGSM, PGD, BIM, SQ, CW, JSMA. При этом состязательное обучение модели в экспериментах выполнено на данных, содержащих состязательные примеры только для методов атак FGSM, PGD, BIM. Набор данных, использованный в экспериментах, являлся сбалансированным. Оценена средняя точность распознавания изображений, в условиях отправки на модель изображений, созданных с использованием указанных видов атак. Сделаны выводы, что состязательное обучение эффективно только для противодействия атакам, которые использовались во время обучения модели, а методы предварительной обработки поступающих данных эффективны только против более простых атак. Средняя точность распознавания в случае применения разработанного метода составила 0,94, что существенно выше рассмотренных методов противодействия атакам. Показано, что точность без применения методов противодействия составляет величину около 0,19, а при состязательном обучении — 0,79, пространственном сглаживании — 0,58, сжатии признаков — 0,88, JPEG-сжатии — 0,37, минимизации общей дисперсии — 0,58, оборонительной дистилляция — 0,44. При этом точность распознавания при атаках FGM, PGD, BIM, SQ, CW, JSMA составила соответственно 0,99, 0,99, 0,98, 0,98, 0,99, 0,73. Разработанный метод представляет более универсальное решение по противодействию всем видам атак, а также достаточно эффективно работает при противодействии сложным состязательным атакам, таким как атаки CW и JSMA. **Обсуждение.** Разработанный метод позволяет повысить точность распознавания с применением машинного обучения при атаках уклонения и, в отличие от состязательного обучения, повышает точность распознавания на состязательных данных, создаваемых с применением атак, не использованных при обучении. Полученные результаты полезны исследователям и специалистам в области машинного обучения.

Ключевые слова
методы машинного обучения, состязательные атаки, защитные механизмы, информационные системы на базе искусственного интеллекта, состязательное обучение

Ссылка для цитирования: Воробьева А.А., Матузко М.А., Сивков Д.И., Сафиуллин Р.И., Менщиков А.А. Новый метод противодействия состязательным атакам уклонения на информационные системы, основанные на искусственном интеллекте // Научно-технический вестник информационных технологий, механики и оптики. 2024. Т. 24, № 2. С. 256–266 (на англ. яз.). doi: 10.17586/2226-1494-2024-24-2-256-266

Introduction

Artificial Intelligence (AI) and Machine Learning (ML) methods are constantly being improved and applied in the most diverse areas of modern life. AI-based systems are vulnerable to attacks, so called adversarial attacks [1].

An adversarial attack is a generalized name for attacks on AI systems including methods of deceiving a Neural Network (NN) to change the system "response" to what the attacker needs and disrupt its performance. These attacks can be performed both at the stage of training the model, and at the stage of its operation [2]. They can be carried out on image recognition systems (photo, video, audio) and are implemented using adversarial samples — data samples in which minor perturbations have been introduced, leading to incorrect recognition [3]. Such perturbations can include adding noise or changing several pixels in the image. The important fact is that the distortions are invisible to humans.

For example, in biometric systems, adding noise or pixels to a person's face image can cause the system to misidentify them. This manipulation increases the security risks for information systems and allows attackers to gain unauthorized access. NN are most susceptible to these attacks, but some classical ML methods are also vulnerable, such as the support vector machine.

The relevance of this research is due to the increasing use of information systems powered by AI, and the rise in security risks associated with adversarial attacks on these systems.

The goal of this study is to improve the accuracy of an image recognition model based on a convolutional NN under conditions of adversarial evasion attacks.

The image recognition problem considered in this work is a multi-class classification task where an image must be categorized into three or more classes. In adversarial evasion attack images are perturbed in such a way that the model is unable to correctly classify them. So, the image recognition task in conditions of adversarial attacks is to correctly classify both normal images and adversarial images.

The practical significance of this study lies in the development of a new method for countering adversarial evasion attacks in information systems based on AI. This method, which we refer to as Counter-Evasion Adversarial Attack (CEAA), will help to protect AI-based systems from these attacks.

The research aims to create and integrate a method that can counter adversarial evasive attacks targeting AI-based information systems. It involves the development and theoretical description of a specific algorithm designed for this purpose. This method is then integrated into the AI-based system. Experimental studies are conducted to evaluate the quality and effectiveness of the method as well as to compare it to other state-of-the-art methods.

Related research

The first who discovered the susceptibility of NN to adversarial attacks were Christian Szegedy, Wojciech Zaremba et al. [1]. They proposed a rather controversial explanation for this phenomenon linking it with the extreme nonlinearity of deep NN in combination with insufficient model averaging and insufficient regularization of the controlled learning task. Then Carlini et al. [4] and Zhang et al. [5] independently found vulnerabilities in automatic speech recognition and voice control systems. Kurakin et al. [6] have shown attacks on autonomous vehicles where an adversarial attack manipulates road signs to trick a trained NN. Since Shegedi's discovery, scientists have focused on Adversarial Learning (AL) to improve the security of NN. Also in recent years, various methods of protection against adversarial attacks have been proposed. All the proposed defense mechanisms proved to be effective against certain classes of attacks, but none of them can be used as a universal solution for all types of attacks. In addition, the implementation of protection methods can lead to a decrease in the performance and efficiency of the NN.

In this study, we have considered adversarial attacks on information systems that perform image recognition tasks. These include systems for biometric identification, medical image classification [7], and countering the distribution of illegal content [8].

This paper focuses on the following types of evasion attacks which are the most common due to the ease of their implementation for the attacker (software implementation in many well-known software libraries and low requirements for computing resources):

— Fast Gradient Sign Method Attack (FGSM) [3];
— Square Method Attack (SQ) [9];
— The Projected Gradient Descent Attack (PGD) [10, 11];
— The Basic Iterative Method Attack (BIM) [12];
— Carlini and Wagner Attack (CW) [13];
— Jacobian Saliency Map Attack (JSMA) [14].

Important to note that these attacks have a high success rate, do not require information about the target model, and are resource efficient from the attacker's point of view.

The increasing threat of the adversarial attacks is widely known and described in reports from the IT-companies[1], the government[2], and the intelligence services [2]. However,

[1] IBM, Trustworthy AI [Electronic resource]. Available at: https://research.ibm.com/topics/trustworthy-ai, free. In Russian (accessed: 19.02.2024).

[2] National Cyber Security Centre NCSC, Annual Review 2023 [Electronic resource]. Available at: https://www.ncsc.gov.uk/collection/annual-review-2023/technology/case-study-cyber-security-ai, free. In Russian (accessed: 19.02.2024).

for the moment, most of scientific research focuses on attacks itself, not on the countermeasures.

Depending on the measures taken, it is possible to classify the protection methods into modification of training or input data, models modification, and using auxiliary tools.

Data modification can be performed during model training or when the model is deployed within the system. This does not require any additional configuration of the model or extensive calculations. Methods within this category include AL, portability blocking, data randomization, data transformation, and data compression.

In model modification, changes are made to the original model architecture or model parameters (by adding extra layers or sub-networks, changing the loss or activation function). This does not require modifying the input data or generating Adversarial Examples (AEs) for training, but it does affect the complexity of model training and the architecture of the model. Examples of methods in this group include gradient masking, defensive distillation, feature squeezing, Deep Contract Network, model masking, and the use of Parseval Networks.

Using auxiliary tools helps to keep the original model intact while adding external models to defend against attacks. These techniques are quite effective in the face of black-box and white-box attacks. However, the main limitation of these tools is that they are quite complex to set up and configure. Some examples of such tools include Defense-GAN and MagNet.

An analytical review of relevant papers in this research has allowed us to identify the most effective methods for countering the attacks mentioned above. These include:

— Data modification:
  — adversarial Learning (AL) [15];
  — JPEG Compression (JC) [16];
  — total Variance Minimization (TVM) [17];
  — feature Squeezing with reducing the color bit depth (FS) [18];
  — spatial Smoothing (SS) [18].
— Model modification:
  — defensive Distillation (DD) [19].

One of the promising methods for countering adversarial attacks is AL. The basis of this method is the addition of AEs to the training dataset, which leads to an increase in the model accuracy on adversarial data. This allows the model to correctly classify both original images and adversarial examples. However, there is no way to account for adversarial attacks of unknown types, which limits the effectiveness of AL. The method is only effective against adversarial attacks that were included in the training process. Additionally, it is not resilient to black-box attacks where the attacker creates AEs using a locally trained model.

The main idea behind the JC is that the input data is transformed into a more condensed form which is then passed on to the model for processing. This process aims to preserve the structure of the input data while making it more challenging or impossible for an attacker to attack the model directly. Compression can help reduce the model sensitivity to minor changes in the input, which can be exploited by an adversary to carry out an adversarial attack. The JC has several benefits when it comes to defending against such attacks. By reducing the model reliance on small changes in input data and reducing the amount of available information to an attacker, the attack becomes less effective.

An alternative approach to address adversarial perturbations is the TVM method which uses a compressed sensing technique that combines pixel dropout and minimization of total variation. In this method, a small subset of pixels is randomly selected, and then an image corresponding to those pixels is reconstructed. The resulting image is free of adversarial perturbations. The JC and TVM methods are quite effective against FGSM and SQ adversarial attacks, but they still cannot provide effective protection against more powerful adversarial attacks such as CW attacks.

The main idea behind FS is to simplify the data representation thus reducing the impact of low-sensitivity attacks. If models are trained on the same data but with different levels of FS, the results of their work will be similar. Meanwhile, an AE that works successfully on the original model is unlikely to work on another model. By calculating the pairwise difference between the outputs of the original and additional models, selecting the maximum value from them, and comparing it to a predetermined threshold, it can be concluded that an input example is adversarial. There are two heuristics methods: reducing the color depth, which means encoding the color with fewer values, and using a smooth filter on an image (SS).

SS (also known as blurring) is a set of techniques used in image processing to reduce noise in images or to create a less pixelated output. Smoothing techniques are either local (using nearby pixels to smooth each individual pixel) or non-local (using larger areas instead of nearby pixels). However, the SS method itself has some limitations. This method is not very effective against certain types of attacks, and using it alone to counter adversarial attacks may not result in an acceptable level of model performance when implementing attacks. While FS and SS methods can effectively prevent certain attacks, they can also reduce the accuracy on real-world data.

DD uses two-stage data processing through distillation. Distillation is a training procedure in which a model is trained to predict probabilities obtained from another model that has previously been trained. The advantage of this approach is that it provides a smoother loss function that is more generalizable for an unknown dataset and has higher accuracy even with AEs. However, with the rise of black-box attacks, DD methods can be easily bypassed due to the robustness of AEs against all models.

The developed CEAA method combines both the AL approach and processes the images provided as input to the model in order to reduce the impact of adversarial perturbations on the model.

A new method for countering evasion adversarial attacks on information systems based on artificial intelligence

In general, an information system based on AI has the following components:
— a source of input data;
— an input data processor which prepares the data for transfer to the ML model (this could be any type of NN or "classical" ML algorithms);
— the ML model itself;
— a model output handler.

The input data processor is responsible for cleaning and transforming the raw data into a format that is suitable for the ML model. This could involve removing outliers, normalizing data, or performing other pre-processing steps. Once the data is prepared, it is passed on to the ML model which uses algorithms such as deep learning or statistical models to analyze it. The model then generates predictions or outputs based on the input data. Finally, the model output is processed by the output handler, which may involve further refinement or interpretation of the results. This ensures that the information system provides accurate and reliable output.

The generalized scheme of an information system based on AI is shown in Fig. 1.

The proposed CEAA method is designed to create image recognition models (NN or ML-model) that are resistant to adversarial evasion attacks. It aims to counter these attacks by changing the existing model and adding a data processing unit to the input.

The CEAA method includes two stages: AL of a model (*M*) and preprocessing data supplied to the input of the model. A flowchart of the developed CEAA method is shown in Fig. 2.

At the stage 1 the following steps are performed:
— import of a dataset (*data*) containing original images without adversarial attacks;
— generation of an adversarial dataset (*advData*) by implementing adversarial attacks on the original dataset;
— training the model *M* on an *advData*;
— validation of the model *M* on *advData*;
— saving the model *M*.

The stage 2 involves preprocessing the data that is fed into a model, *M*. This stage is based on techniques for modifying the data to protect against adversarial attacks. Specifically, it involves transforming an image in order to reduce the impact of adversarial perturbations on the classification outcome. This process involves the following steps:
1) obtaining an image (*sample*);
2) *sample* transformation to obtain *sampleB* using:
   a) feature squeezing method;
   b) JPEG compression method;
3) transferring the *sampleB* to the input of the model *M* prepared at the stage 1;
4) recognition of *sampleB*, that is equivalent of the input *sample*, with model *M*.

The scientific novelty of the method is characterized by the original combination of methods for countering adversarial attacks: AL of a model and data transformation.

Integration of the developed method for countering adversarial evasion attacks with information systems based on artificial intelligence

Thus, the method consists of the following blocks: input data processing and training resistant ML model on AEs.

The input data preprocessing stage performs image modification functions in order to reduce the effect of adversarial perturbations on the model. The model trained on AEs performs the function of classifying input data.

After embedding the proposed method for counteracting adversarial attacks, the general block diagram of a data-driven AI-based information system will take the form shown in Fig. 3.

Thus, after integrating the proposed method of countering adversarial evasion attacks into the information system, the system would operate in the following way:
— collection of input data from sensors or data stores;
— transformation of input data to ensure correct work with the model;
— transformation of the image to minimize the impact of adversarial distortions on model operation;
— processing of the transformed data using the model to generate the output;
— taking action based on the output from the model.

A generalized scheme of the AI-based system, after the integration of CEAA, is shown in Fig. 4.

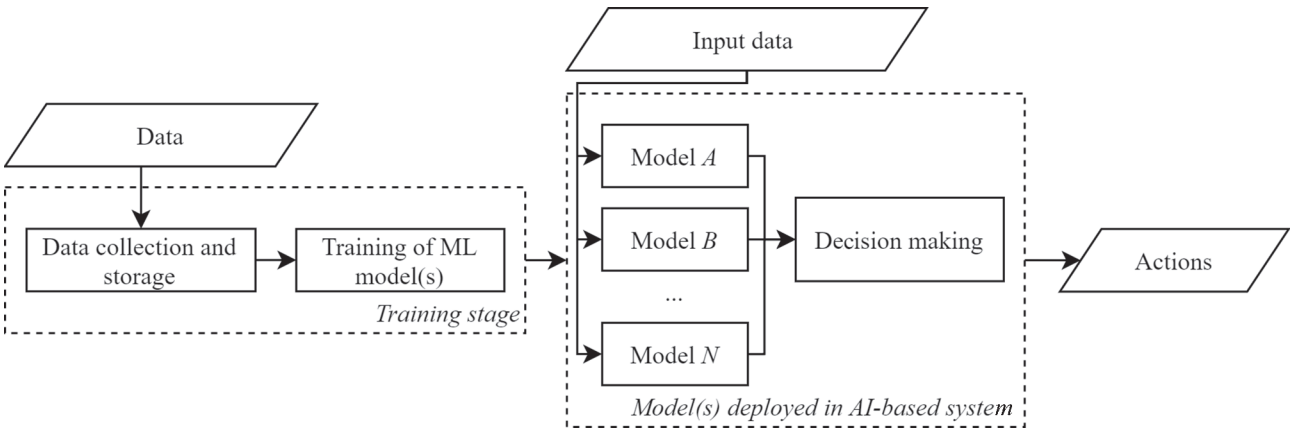

Fig. 1. Generalized scheme of AI-based system

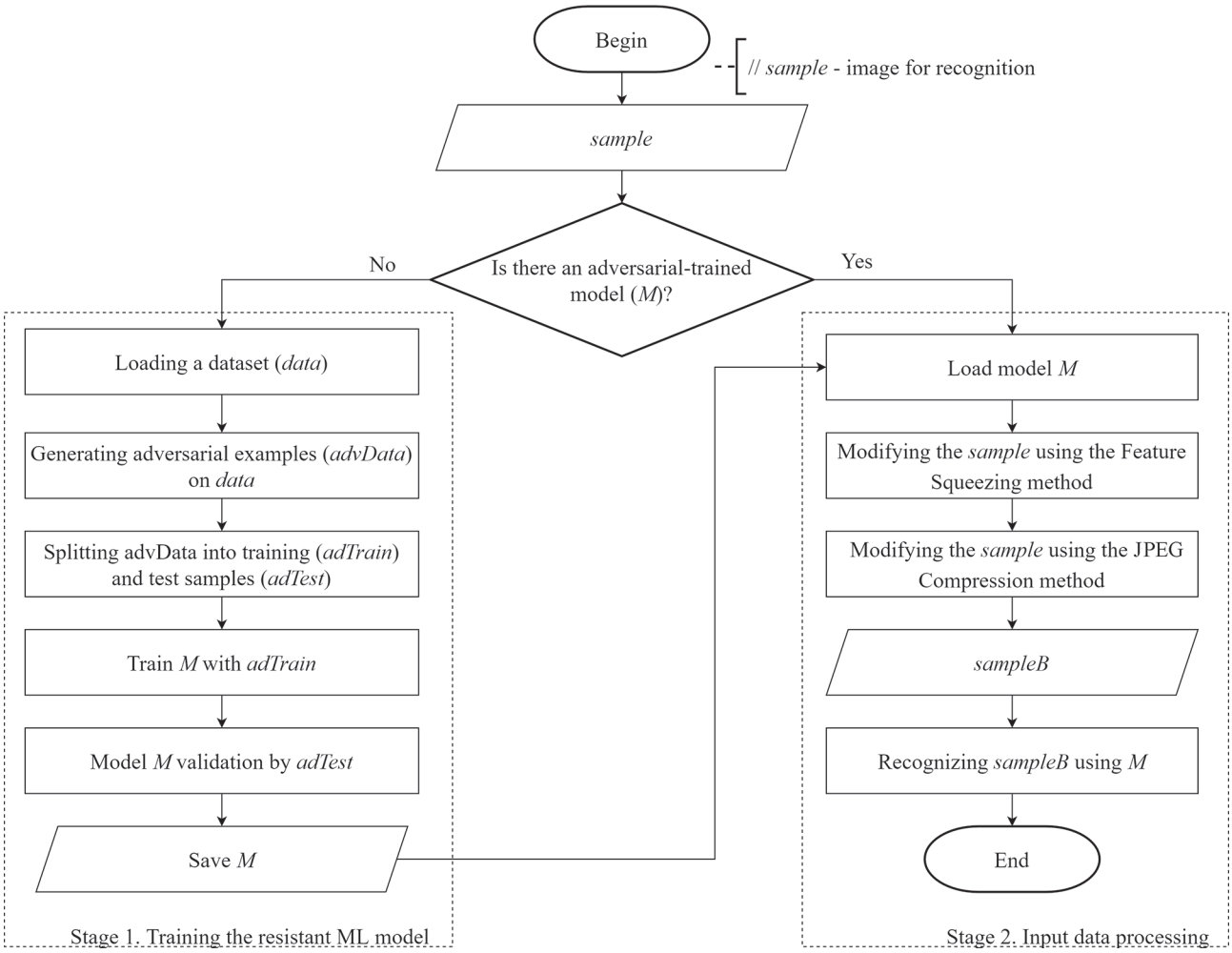

Fig. 2. Flowchart of the developed method for countering adversarial evasion attacks on information systems based on AI

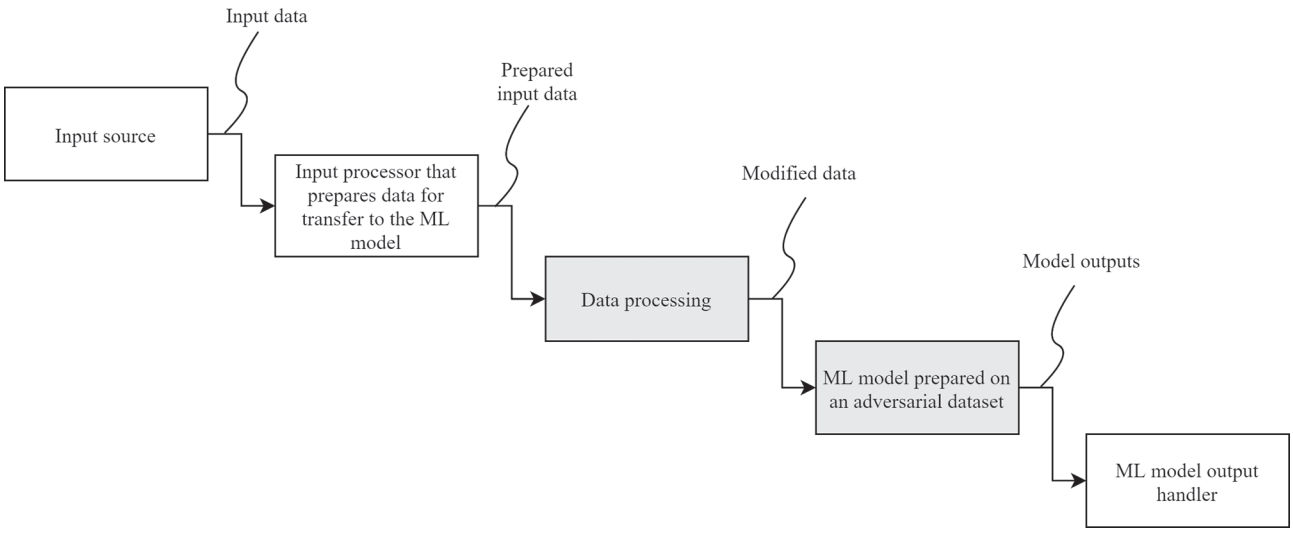

Fig. 3. Generalized block diagram of an information system based on AI, with the integration of the proposed method for countering adversarial evasion attacks

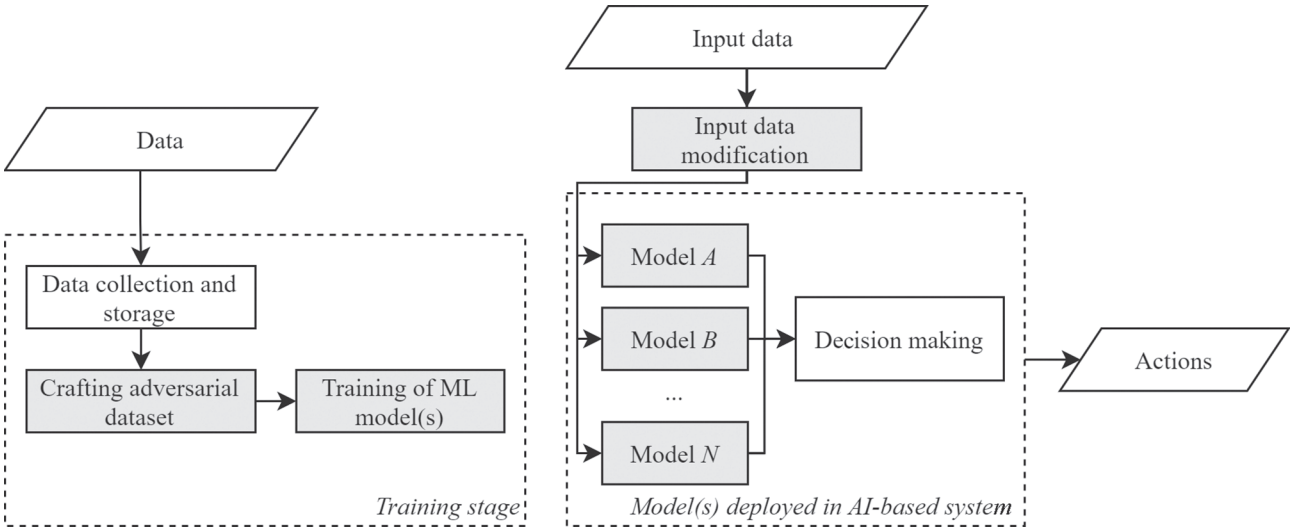

Fig. 4. Generalized scheme of AI-based system after integration of CEAA

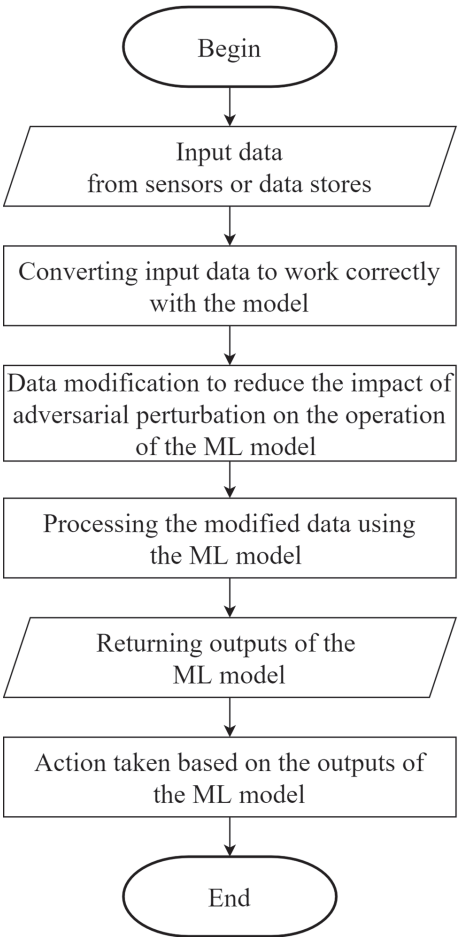

Fig. 5. Flowchart of the information system after the integration of the proposed method

The algorithm of the information system after the integration of the proposed method is shown in Fig. 5.

Experimental studies to assess the quality of the developed method for countering adversarial evasion attacks on AI-based systems

There are two main objectives of the experimental research. The first is to evaluate the impact of the proposed CEAA method on the accuracy, precision, and recall of image recognition using a ML model (Convolutional Neural Network, CNN) under adversarial evasion attacks. The second is to assess the effectiveness of CEAA compared to other existing methods for countering adversarial attacks.

Experimental setup

As mentioned above, the image recognition task that is considered in this paper is a multiclass classification task. In order to evaluate the performance of a classification model, it is common practice to use the following metrics: accuracy, precision, and recall.

The accuracy is calculated using the formula

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \quad (1)$$

The calculation of precision is made according to the formula

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \quad (2)$$

The recall is calculated using the formula

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \quad (3)$$

Where TP (True Positive) — correctly classified objects of a positive class, TN (True Negative) — correctly classified objects of a negative class, FP (False Positive) — incorrectly classified objects classified by the classifier as positive, FN (False Negative) — incorrectly classified objects, classified by the classifier as negative.

For experimental purposes, a CNN with the architecture shown in Table 1 was developed. The selection of this type of NN is based on its high accuracy in recognizing and classifying images, as well as its smaller number of

Table 1. Architecture of the CNN model used in experiments

Layer	Output shape	Activation function
Conv2D	(None, 26, 26, 32)	relu
MaxPooling2D	(None, 13, 13, 32)	—
Conv2D	(None, 11, 11, 64)	relu
MaxPooling2	(None, 5, 5, 64)	—
Flatten	(None, 1600)	—
Dropout	(None, 1600)	—
Dense	(None, 10)	softmax

adjustable parameters and its resistance to rotation and translation of the recognized images. The choice of this particular architecture is justified by its high accuracy rates on test data, with only a small number of layers in the network.

During the preparation of the experiments, the following requirements were formed for the dataset: all images must be square and have the same size in pixels in order for them to work correctly with the NN; the data must be labeled; the minimum number of images for one class is 500, and it should allow us to assess the accuracy of image recognition in conditions of adversarial attacks, before and after applying the CEAA method.

To train the model, the Modified National Institute of Standards and Technology (MNIST) image dataset was chosen as it is the most efficient in terms of model preparation and is widely used by the scientific community for experimental evaluation of protection methods against adversarial attacks.

The dataset was balanced, meaning that each class has approximately the same number of training and testing samples as the other classes. There are 10 classes in the dataset, with approximately 1,000 images per class, resulting in a total of 10,000 samples. In all of the experiments, the dataset was divided into training and testing sets in a ratio of 80:20.

At first the CNN was trained on a prepared dataset that did not contain adversarial attacks. The following parameters were used: batch size — 128, number of epochs — 15. The model was given the following name: Image Recognition Model (IRM). The performance of IRM was then evaluated according to formula (1), the accuracy obtained was 0.9922.

Then based on the prepared dataset and the IRM model, adversarial examples for FGSM, SQ, PGD, BIM, CW and JSMA attacks were crafted, six independent adversarial datasets were created.

After that to implement the AL stage of the developed CEAA method, a second model was trained on data containing only AEs for FGSM, PGD and BIM attacks. The model was not trained on all considered attacks for the purpose of evaluating the developed CEAA method objectively. The performance of this model was evaluated according to formula (1), the accuracy obtained was 0.9911.

Evaluation of the quality of the image recognition model under FGSM, SQ, PGD, BIM evasion attacks before and after implementation of the developed method

The aim of the experimental research is to determine the accuracy, precision and recall of the image recognition model, both before and after implementing the CEAA method.

In two series of experiments the accuracy, precision, and recall of image recognition were assessed using formulas (1)–(3):

1) for the base IRM without any evasion attacks countermeasures on normal images and adversarial images;
2) for the image recognition model with the implementation of the developed CEAA method on normal images and adversarial images.

The inputs to the base IRM model and the model prepared with CEAA were fed with images from the generated datasets including images that had been subjected to adversarial attacks, such as FGSM, PGD, and BIM. The results of these experiments are presented in Table 2.

The results of the first set of experiments demonstrate the vulnerability of the original model to adversarial attacks. Although the model performance indicators on the initial dataset are high, it would be easy for an attacker to "trick" such a model with adversarial examples.

In the second set of experiments, the accuracy, precision, and recall of the IRM after implementing of the developed CEAA method were evaluated. The results of these experiments showed that the values of the performance metrics of the image recognition model slightly decreased on the initial data after applying the CEAA method. But at the same time, they increased significantly in conditions of the implemented adversarial attacks. It is worth noting that the model performance

Table 2. Comparison of accuracy, precision and recall of the image recognition using CNN under FGSM, SQ, PGD, BIM evasion attacks before and after implementation of the developed method

Attack method	Accuracy		Precision		Recall	
	Base IRM	CEAA	Base IRM	CEAA	Base IRM	CEAA
No attack	0.9922	0.9887	0.9921	0.9886	0.9921	0.9886
FGSM	0.3754	0.9854	0.4563	0.9853	0.3764	0.9852
PGD	0.0735	0.9862	0.0811	0.9861	0.0747	0.9861
BIM	0.0671	0.9824	0.0671	0.9822	0.0686	0.9824
SQ	0.1695	0.9800	0.2155	0.9798	0.1721	0.9798
Average for all attacks	0.1714	0.9835	0.2050	0.9834	0.1730	0.9834

Table 3. Comparison of the image recognition accuracy before and after implementation of the developed CEAA method with other existing defense methods for various attack types

Attack method	Adversarial attacks defense method							
	IRM (NM)	**CEAA**	AL	SS	FS	JC	TVM	DD
	Accuracy							
NA	0.9922	**0.9887**	0.9911	0.9686	0.9896	0.9921	0.7952	0.9916
FGSM	0.3754	**0.9854**	0.9871	0.7548	0.9659	0.4478	0.6508	0.5268
PGD	0.0735	**0.9862**	0.9872	0.1649	0.9284	0.0736	0.4772	0.0735
BIM	0.0671	**0.9824**	0.9827	0.1589	0.8301	0.0671	0.4749	0.0671
SQ	0.1695	**0.9800**	0.7992	0.5637	0.9766	0.4934	0.4021	0.6527
CW	0.4417	**0.9873**	0.9876	0.8930	0.8538	0.5651	0.7503	0.7406
JSMA	0.0011	**0.7284**	0.0010	0.9290	0.7410	0.5669	0.7175	0.5812
Average for all attacks	0.1881	**0.9416**	0.7908	0.5774	0.8826	0.3690	0.5788	0.4403

NOTE. IRM (NM) — without using any adversarial attack countermeasures, NA — normal images (without attacks).

remained acceptable even in the presence of adversarial attacks, thus proving the effectiveness of the developed CEAA method in counteracting adversarial attacks.

Evaluation the effectiveness of the developed method of countering adversarial evasion attacks in comparison with other existing methods

The aim of this experimental study is to compare the developed CEAA method with the existing methods in terms of accuracy.

In the series of experiments, the image recognition accuracy was evaluated using formula (1) on adversarial data, crafted with FGSM, SQ, PGD, BIM, CW, JSMA attacks. Developed CEAA method was compared on accuracy with various methods of countering adversarial attacks (AL, SS, FS, JC, TVM, DD).

Normal and adversarial images were fed to the model input:

— for the base IRM without any evasion attacks countermeasures;
— for the image recognition model with the implementation of the developed CEAA method;
— for the image recognition model with the AL, SS, FS, JC, TVM, DD countermeasures;
— the results of the series of experiments are presented in Table 3 and in Fig. 6.

It is worth noting the differences in indicators between the CEAA method and the AL method. There is a slight difference in accuracy between CEAA and AL for attacks FGSM, PGD, BIT that were used to train the model in AL method. However, at the same time, the developed CEAA method produces much better results for attacks that were not included during the AL, which can be clearly seen in attacks such as SQ and JSMA. The accuracy for the CW attack is approximately the same.

Average image recognition accuracy of the adversarial data without any defense is around 0.19, and after implementation of the developed method — 0.94. Method performs better than the existing methods, accuracy of the image recognition model only with Adversarial Learning is 0.79, Spatial Smoothing — 0.58, Feature Squeezing — 0.88, JPEG compression — 0.37, Total Variance Minimization — 0.58, Defensive Distillation — 0.44.

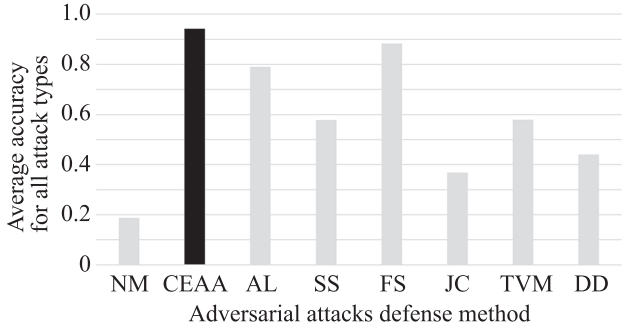

Fig. 6. Average image recognition accuracy on adversarial data for various countermeasures

The experimental results indicate that the proposed CEAA method is more effective in countering adversarial evasion attacks than other methods analyzed. It provides better image recognition accuracy compared to existing methods. Therefore, the developed CEAA approach allows for high performance indicators of a model, even when adversarial attacks are present, which were not taken into account when creating the adversarial model.

Conclusion and future work

In conclusion, the study on the development of a method to counter adversarial evasion attacks in AI-based information systems has shown promising results. When using the novel method, there was a slight decrease in model performance on initial data, but it significantly improved resilience and accuracy against adversarial attacks. Interestingly, performance remains acceptable even under attacks highlighting the effectiveness of the method.

Comparative experiments also revealed that this method outperformed existing techniques, especially against novel adversarial attacks not considered during model training. The significant improvement in model performance against such unexpected attacks demonstrates the method robustness and adaptability.

The developed method can counteract adversarial evasion attacks. The novelty of the solution lies in the

combination of adversarial learning and the preprocessing of input data for the model. This approach has practical value in improving the accuracy of the model under the impact of adversarial attacks.

Future research will focus on optimizing information systems based on artificial intelligence in order to enhance resilience against a wider range of attacks, while maintaining performance and enhancing real-time defensive capabilities, as well as ensuring adaptability to different models.

References

1. 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
2. Tabassi E., Burns K.J., Hadjimichael M., Molina-Markham A.D., Sexton J.T. *A taxonomy and terminology of adversarial machine learning: NIST IR*, 2019, pp. 1–29.
3. Goodfellow I.J., Shlens J., Szegedy C. Explaining and harnessing adversarial examples. *arXiv*, 2015, arXiv:1412.6572. https://doi.org/10.48550/arXiv.1412.6572
4. Carlini N., Mishra P., Vaidya T., Zhang Y., Sherr M., Shields C., Wagner D., Zhou W. Hidden voice commands. *Proc. of the 25th USENIX Security Symposium*, 2016, pp. 513–530.
5. Zhang G., Yan C., Ji X., Zhang T., Zhang T., Xu W. Dolphinattack: Inaudible voice commands. *Proc. of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 103–117. https://doi.org/10.1145/3133956.3134052
6. Kurakin A., Goodfellow I.J., Bengio S. Adversarial machine learning at scale. *International Conference on Learning Representations (ICLR)*, 2017.
7. Li X., Zhu D. Robust detection of adversarial attacks on medical images. *Proc. of the 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 2020, pp. 1154–1158. https://doi.org/10.1109/isbi45749.2020.9098628
8. Imam N.H., Vassilakis V.G. A survey of attacks against twitter spam detectors in an adversarial environment. *Robotics*, 2019, vol. 8, no. 3, pp. 50. https://doi.org/10.3390/robotics8030050
9. Andriushchenko M., Croce F., Flammarion N., Hein M. Square attack: a query-efficient black-box adversarial attack via random search. *Lecture Notes in Computer Science*, 2020, vol. 12368, pp. 484–501. https://doi.org/10.1007/978-3-030-58592-1_29
10. Deng Y., Karam L.J. Universal adversarial attack via enhanced projected gradient descent. *Proc. of the 2020 IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 1241–1245. https://doi.org/10.1109/icip40778.2020.9191288
11. Madry A., Makelov A., Schmidt L., Tsipras D., Vladu A. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations (ICLR)*, 2018.
12. Kurakin A., Goodfellow I.J., Bengio S. Adversarial examples in the physical world. *Artificial Intelligence Safety and Security*, 2018, pp. 99–112. https://doi.org/10.1201/9781351251389-8
13. Carlini N., Wagner D. Towards evaluating the robustness of neural networks. *Proc. of the 2017 IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 39–57. https://doi.org/10.1109/sp.2017.49
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. Lowd D., Meek C. Adversarial learning. *Proc. of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, 2005, pp. 641–647. https://doi.org/10.1145/1081870.1081950
16. Das N., Shanbhogue M., Chen S.-T., Hohman F., Chen L., Kounavis M.E., Chau D.H. Keeping the bad guys out: Protecting and vaccinating deep learning with JPEG compression. *arXiv*, 2017, arXiv:1705.02900. https://doi.org/10.48550/arXiv.1705.02900
17. Guo C., Rana M., Cisse M., van der Maaten L. Countering adversarial images using input transformations. *International Conference on Learning Representations (ICLR)*, 2018.
18. Xu W., Evans D., Qi Y. Feature squeezing: detecting adversarial examples in deep neural networks. *Proc. of the 2018 Network and Distributed System Security Symposium*, 2018. https://doi.org/10.14722/ndss.2018.23198

Литература

1. 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
2. Tabassi E., Burns K.J., Hadjimichael M., Molina-Markham A.D., Sexton J.T. A taxonomy and terminology of adversarial machine learning: NIST IR. 2019. P. 1–29.
3. Goodfellow I.J., Shlens J., Szegedy C. Explaining and harnessing adversarial examples // arXiv. 2015. arXiv:1412.6572. https://doi.org/10.48550/arXiv.1412.6572
4. Carlini N., Mishra P., Vaidya T., Zhang Y., Sherr M., Shields C., Wagner D., Zhou W. Hidden voice commands // Proc. of the 25th USENIX Security Symposium. 2016. P. 513–530.
5. Zhang G., Yan C., Ji X., Zhang T., Zhang T., Xu W. Dolphinattack: Inaudible voice commands // Proc. of the 2017 ACM SIGSAC Conference on Computer and Communications Security. 2017. P. 103–117. https://doi.org/10.1145/3133956.3134052
6. Kurakin A., Goodfellow I.J., Bengio S. Adversarial machine learning at scale // International Conference on Learning Representations (ICLR). 2017.
7. Li X., Zhu D. Robust detection of adversarial attacks on medical images // Proc. of the 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). 2020. P. 1154–1158. https://doi.org/10.1109/isbi45749.2020.9098628
8. Imam N.H., Vassilakis V.G. A survey of attacks against twitter spam detectors in an adversarial environment // Robotics. 2019. V. 8. N 3. P. 50. https://doi.org/10.3390/robotics8030050
9. Andriushchenko M., Croce F., Flammarion N., Hein M. Square attack: a query-efficient black-box adversarial attack via random search // Lecture Notes in Computer Science. 2020. V. 12368. P. 484–501. https://doi.org/10.1007/978-3-030-58592-1_29
10. Deng Y., Karam L.J. Universal adversarial attack via enhanced projected gradient descent // Proc. of the 2020 IEEE International Conference on Image Processing (ICIP). 2020. P. 1241–1245. https://doi.org/10.1109/icip40778.2020.9191288
11. Madry A., Makelov A., Schmidt L., Tsipras D., Vladu A. Towards deep learning models resistant to adversarial attacks // International Conference on Learning Representations (ICLR). 2018.
12. Kurakin A., Goodfellow I.J., Bengio S. Adversarial examples in the physical world // Artificial Intelligence Safety and Security. 2018. P. 99–112. https://doi.org/10.1201/9781351251389-8
13. Carlini N., Wagner D. Towards evaluating the robustness of neural networks // Proc. of the 2017 IEEE Symposium on Security and Privacy (SP). 2017. P. 39–57. https://doi.org/10.1109/sp.2017.49
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. Lowd D., Meek C. Adversarial learning // Proc. of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. 2005. P. 641–647. https://doi.org/10.1145/1081870.1081950
16. Das N., Shanbhogue M., Chen S.-T., Hohman F., Chen L., Kounavis M.E., Chau D.H. Keeping the bad guys out: Protecting and vaccinating deep learning with JPEG compression // arXiv. 2017. arXiv:1705.02900. https://doi.org/10.48550/arXiv.1705.02900
17. Guo C., Rana M., Cisse M., van der Maaten L. Countering adversarial images using input transformations // International Conference on Learning Representations (ICLR). 2018.
18. Xu W., Evans D., Qi Y. Feature squeezing: detecting adversarial examples in deep neural networks // Proc. of the 2018 Network and Distributed System Security Symposium. 2018. https://doi.org/10.14722/ndss.2018.23198

19. Papernot N., McDaniel P., Wu X., Jha S., Swami A. Distillation as a defense to adversarial perturbations against deep neural networks. *Proc. of the 2016 IEEE Symposium on Security and Privacy (SP)*, 2016, pp. 582–597. https://doi.org/10.1109/sp.2016.41

19. Papernot N., McDaniel P., Wu X., Jha S., Swami A. Distillation as a defense to adversarial perturbations against deep neural networks // Proc. of the 2016 IEEE Symposium on Security and Privacy (SP). 2016. P. 582–597. https://doi.org/10.1109/sp.2016.41

Authors

Alisa A. Vorobeva — PhD, Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, sc 57191359167, https://orcid.org/0000-0001-6691-6167, vorobeva@itmo.ru

Maxim A. Matuzko — Engineer, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0009-0006-2179-6847, mxmmtzk@gmail.com

Dmitry I. Sivkov — Engineer, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0000-0002-3008-3789, 485280@mail.ru

Roman I. Safiullin — Engineer, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0009-0004-8635-9432, romsaaf@mail.ru

Alexander A. Menshchikov — PhD, Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, sc 57196052901, https://orcid.org/0000-0002-2287-4310, menshikov@itmo.ru

Авторы

Воробьева Алиса Андреевна — кандидат технических наук, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, sc 57191359167, https://orcid.org/0000-0001-6691-6167, vorobeva@itmo.ru

Матузко Максим Александрович — инженер, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0009-0006-2179-6847, mxmmtzk@gmail.com

Сивков Дмитрий Игоревич — инженер, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0000-0002-3008-3789, 485280@mail.ru

Сафиуллин Роман Ильшатович — инженер, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0009-0004-8635-9432, romsaaf@mail.ru

Менщиков Александр Алексеевич — кандидат технических наук, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, sc 57196052901, https://orcid.org/0000-0002-2287-4310, menshikov@itmo.ru

Received 06.02.2024
Approved after reviewing 02.03.2024
Accepted 27.03.2024

Статья поступила в редакцию 06.02.2024
Одобрена после рецензирования 02.03.2024
Принята к печати 27.03.2024