DOI: 10.15514/ISPRAS-2023-35(2)-3

Investigation of Adversarial Attacks on Pattern Recognition Neural Networks

D.V. Kotlyarov. ORCID: 0009-0003-2530-4043<den14kotlyarov@yandex.ru>
G.D. Dyudyun, ORCID: 0009-0008-1256-0204 <gleb.dudun@gmail.com>
N.V. Rzhevskaya, ORCID: 0009-0002-1285-4196 <natalia070901@gmail.com>
M.A. Lapina, ORCID: 0000-0001-8117-9142 <mlapina@ncfu.ru>
M.G. Babenko, ORCID: 0000-0001-7066-0061 <mgbabenko@ncfu.ru>

North-Caucasus Federal University,
1, Pushkin Street, Stavropol. 355017, Russia

Abstract. This article discusses the algorithm for creating a neural network based on pattern recognition. Several types of attacks on neural networks are considered, the main features of such attacks are described. An analysis of the Adversarial attack was carried out. The results of experimental testing of the proposed attack are presented. Confirmation of the hypothesis about the decrease in the accuracy of recognition of the neural network during the implementation of the attack by an attacker was obtained.

Keywords: neural network; machine learning; pattern recognition; artificial intelligence; attack algorithm; information security; Adversarial attack, malicious machine learning

For citation: Kotlyarov D.V., Dyudyun G.D., Rzhevskaya N.V., Lapina M.A., Babenko M.G. Investigation of Adversarial Attacks on Pattern Recognition Neural Networks. Trudy ISP RAN/Proc. ISP RAS, vol. 35, issue 2, 2023. pp. 35-48. DOI: 10.15514/ISPRAS-2023-35(2)-3

Acknowledgements. This work was supported by the Russian Science Foundation 19-71-10033, https://rscf.ru/project/19-71-10033/

Исследование состязательных атак на нейронные сети распознавания образов

Д. В. Котляров. ORCID: 0009-0003-2530-4043<den14kotlyarov@yandex.ru>
Г. Д. Дюдюн. ORCID: 0009-0008-1256-0204 <gleb.dudun@gmail.com>
Н.В. Ржевская. ORCID: 0009-0002-1285-4196 <natalia070901@gmail.com>
М.А. Лапина. ORCID: 0000-0001-8117-9142 <mlapina@ncfu.ru>
М.Г. Бабенко. ORCID: 0000-0001-7066-0061 <mgbabenko@ncfu.ru>

Северо-Кавказский федеральный университет,
Россия, 355017, г. Ставрополь, ул. Пушкина, 1

Аннотация. В данной статье рассмотрен алгоритм создания нейросети, базирующую на распознавание образов. Рассмотрены несколько видов атак на нейронные сети, описаны основные особенности таких атак. Проведен анализ Adversarial attack. Приводятся результаты экспериментальной апробации предложенной атаки. Получено подтверждение гипотезы о снижении точности распознавания нейросети при реализации атаки злоумышленником.

Ключевые слова: нейронная сеть; машинное обучение; распознавание образов; искусственный интеллект; алгоритм атак; информационная безопасность; состязательная атака; вредоносное машинное обучение

Для цитирования: Котляров Д.В., Дюдюн Г.Д., Ржевская Н.В., Лапина М.А., Бабенко М.Г. Исследование состязательных атак на нейронные сети распознавания образов. Труды ИСП РАН, том 35, вып. 2, 2023 г., стр. 35-48. DOI: 10.15514/ISPRAS–2023–35(2)–3

Благодарности. Работа выполнена при поддержке РНФ 19-71-10033, https://rscf.ru/project/19-71-10033/

1. Introduction

In modern times, such concepts as “neural network”, “artificial intelligence” and other end-to-end technologies are used in various fields, such technologies are tightly integrated into our lives and are often used everywhere. Now nobody is surprised using search algorithms from various advanced companies. Even the older generation now often uses voice assistants, without thinking that these are yet another “brainchild” of a neural network. In fact, a neural network can now do quite a lot: generate scientific reports, write poetry or a song, draw a picture that is not much different from the real one, the main thing is to train it correctly.

But such a network can serve both good and evil, depending on the purpose of the developer. So, for example, when recognizing images, an attacker can purposefully introduce errors into the recognition process, trying to force the system to incorrectly recognize the image being processed [1]. As a result, so-called spoofing attacks appear. Often, such attacks can be used in cases where an attacker seeks to disguise himself as another person and thereby commit illegal actions.

2. State-of-the-art

At present, it is difficult to unambiguously define neural networks. After analyzing the study of several authors, we can say that a neural network or Artificial Neural Networks (ANN) is a learning system, which is a certain mathematical model built on the principle of human neurons, as well as its software implementation [2, 3]. Ivanyuk V.A. claims that artificial neural networks can be used to create intelligent decision-making systems, simulation modeling, expert systems [2].

A neural network is a mathematical model made up of interconnected nodes that work together to solve a problem. The nodes are arranged in layers, and each node performs a simple mathematical operation on the input to produce the output [4].

Kachagina K.S. in her research, gives a range of applications of a neural network in everyday life. So, in this study, we can highlight that the ANN is already used in security organizations, law enforcement agencies, at various factories and much more [3]. Therefore, it is very important to organize the security of these systems, since further the scope of neural systems will only grow.

It should be noted that the main tasks of neural networks are reduced to:

- Classification, that is, the separation of a certain object with a certain attribute from others.
- Prediction, this task often serves the interests of the financial world.
- Recognition, which will help to simplify the work, for example, for law enforcement agencies.
- Solving problems without a teacher.

In recent years, there has been an introduction into information and telecommunication systems as a means of identification, and often authentication of users [1]. According to experts, the introduction of such technologies often brings with it massive discontent from the outside. The reason for this was that the ANN is imperfect. Such a system has several vulnerabilities that will be used to disable it [3].

Since each person is unique, by spoofing biometric data, attackers can describe such data mathematically and use it as input to machine learning algorithms to automate the recognition process, and then use such ANN to replace their identity.

There are many attacks on ANN that prevent the system from working properly. An attacker can carry out large-scale attacks without being noticed. For example, in biometric systems, an attacker can intentionally introduce errors into the process of recognizing biometric data. Ensuring the security of such systems is an important issue.

Article [5] describes how adversarial attacks work by exploiting vulnerabilities in neural networks that can be easily fooled by small noises or modifications to the input data that are imperceptible to humans but can cause the network to misclassify the input data.

Consider some types of attacks on biometric systems that disrupt the recognition process [1]:

- Fast Gradient Sign Method – an attack with noise overlay on the image with each new iteration. This attack is quite effective when constantly analyzing the image. This type of attack is practically unrealizable in the absence of direct access to data.
- Using Infrared LEDs to Change Human Facial Features.
- Overlaying black or white stickers on the image for incorrect recognition.
- The use of devices that allow you to identify a person for another.
- Hostile attacks are a growing concern in the field of artificial intelligence because they can be used to trick neural networks into misclassifying inputs.

One of the first studies on Adversarial Attacks was carried out by Goodfellow et al. [6], who showed that neural networks can be fooled by small noise inputs. Since then, a large amount of research has been done on this topic, including the development of new attack methods and defense strategies.

One of the most common types of contention attacks is the Fast Gradient Sign Method (FGSM), which was introduced in [6]. This method involves calculating the gradient of the loss function with respect to the input data and then modifying the data in the direction of the gradient to maximize the loss. Many subsequent studies have relied on this method, including the iterative FGSM (IFGSM) attack presented by Kurakin et al [7].

Article [5] also explains various types of hostile attacks, such as targeted and non-targeted attacks, and provides examples of real-life applications of hostile attacks, such as manipulating the systems of unmanned vehicles.

Other types of attacks that have been developed include the Jacobian-based saliency map attack [8], which uses the Jacobian matrix to determine the most sensitive input features, and the deep fool attack [9], which generates small perturbations, which minimize the distance between the original input and the misclassified output.

Various strategies have been proposed to protect against these attacks. Adversarial learning involves augmenting the training data with adversarial examples to make the neural network more robust [10], while defensive distillation involves training a separate network to detect adversarial examples [11]. Other protection strategies include randomization, input transformation, and gradient masking [12].

Despite these defense strategies, adversary attacks remain a major threat to machine learning systems. As noted by Akhtar and Mian [13], attacks by the adversary can have serious consequences in the real world, such as causing self-driving cars to misinterpret road signs or medical systems to misdiagnose diseases.

In addition, article [5] discusses some of the techniques that have been developed to defend against adversary attacks, including adversary training and defensive distillation.

In recent years, researchers have also studied the impact of adversarial attacks on object detection systems [14], semantic segmentation [15], and generative models [16]. These studies have shown that enemy attacks are effective against these systems and have proposed new defense methods to improve their reliability.

Despite significant progress in the development of adversarial attacks and defenses, there are still open issues that require further research. One of these tasks is the development of effective and reliable methods of protection. Another challenge is understanding the vulnerabilities of deep learning models to attack by malicious actors and finding ways to fix them.

Several studies [6, 14, 15] have examined the effectiveness of adversarial attacks on various types of Machines Learning (ML) models, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and autoencoders. For example, Xu et al. [17] have shown that adversarial attacks can be effective against RNN-based text classifiers even if the attacks are generated using a different language model.

To protect against enemy attacks, researchers have proposed various defense strategies. One common defense strategy is adversarial learning, which involves training a model on adversarial examples in addition to regular training data [10]. Other security strategies include gradient masking, which involves hiding gradients from an attacker [11], and feature compression, which involves preprocessing input data to remove redundant features [18].

Several recent studies have also explored the use of generative models such as Generative Adversarial Networks (GANs) to generate adversarial examples [14]. These models can be used to create more realistic examples of competitive actions that are harder to detect and more difficult to defend against.

Biometric authentication systems, which use physiological or behavioral characteristics to verify people's identities, have become increasingly popular in recent years. However, these systems are not immune from attacks, and neural networks are used to carry out attacks on biometric data [18].

One common type of attack on biometrics is presentation attacks, also known as spoofing attacks, where an attacker uses a fake or artificial biometric to impersonate a legitimate user. Neural networks have been used to create attacks with a realistic representation of various biometric modalities, including fingerprints, face recognition, and voice recognition.

For example, Wang et al. [19] used a convolutional neural network (CNN) to generate realistic fingerprints that can be used to fake fingerprint recognition systems. Similarly, Nguyen et al. [20] used a generative adversarial network (GAN) to generate synthetic facial images that can be used to fake facial recognition systems.

Other research has focused on using neural networks to launch attacks on biometric data by exploiting vulnerabilities in the biometric system. For example, Li et al. [21] proposed a method for generating adversarial examples for fingerprint recognition systems by distorting the input fingerprint image using a gradient-based optimization method. The resulting fingerprint of the attacker can be used to avoid detection by the biometric system.

To protect against attacks on biometric data, researchers have proposed various defense strategies, including the use of liveness detection techniques to detect attacks on presentations and the use of deep neural networks to increase the resilience of biometric systems to attacks. For example, Tan et al. [22] proposed a deep neural network liveliness detection method for detecting presentational attacks in face recognition systems.

3. Materials and research methods

Using the MNIST dataset as an example, we can consider the principle of building a certain neural network, and then explore its vulnerability. Based on the research of V.A. Ivanyuk, any neural network is mathematically a superposition of regression functions that describe the relationship between the values of the inputs and outputs of the network [2].

The input layer receives data in the form of features or input variables, which are then passed to the first hidden layer. Each node in the hidden layer performs a linear transformation of the input using weights and biases and then applies a non-linear activation function to produce a non-linear output. This output is then fed as input to the next level, and the process is repeated until the final output is obtained.

Starting the study, it is worth saying that the basic element of such systems is the so-called neuron. It is necessary to create a programming model. A neuron in an ANN is an artificial analogue of a real neuron, only represented as a simple mathematical function that determines the rules for generating an output signal based on input data [2]. In the work of K.S. Kachagina stated that a neuron is an imaginary black object with several input and one output hole [3].

A neural network can be represented mathematically as a function $f(x;\ \theta)$, where x is the input, θ are the network parameters (weights and biases), and $f(x;\ \theta)$ is the output of the network.

The weights and biases in the neural network are adjusted during training to optimize network performance. This is done by minimizing the cost function, which measures the difference between projected output and actual output. The backpropagation algorithm is used to update the weights and biases in such a way as to reduce the loss function.

A neural network consists of layers of interconnected nodes, and the calculation of the output of each node can be represented mathematically as (1):

$$z = w \cdot x + b \tag{1}$$

where z – weighted sum of input parameters x, w – scale vector, b – displacement vector.

The output of a node is calculated by applying a non-linear activation function to a weighted sum, which can be represented mathematically according to (2):

$$a = g(z) \tag{2}$$

where $g(x)$ is the activation function.

The node layer output is calculated as follows (3):

$$a^1 = g(z^1) \tag{3}$$

where a^1 – exit of the first layer, z^1– weighted sum of first layer inputs, g is the activation function.

The output of the last layer is the output of the neural network, which can be used for prediction or classification.

During training, the weights and biases of the neural network are adjusted to minimize the cost function, which measures the difference between the predicted output and the actual output. This is done using gradient descent, where the gradient of the cost function with respect to weights and biases is computed, and the weights and biases are updated accordingly.

The backpropagation algorithm is used to efficiently compute the gradient of the cost function with respect to the weights and biases of the network.

Thus, the mathematical logic of a neural network includes calculating the weighted sum of the input data for each node, applying a non-linear activation function, and propagating the output through the layers of the network to obtain the final output. The network weights and biases are optimized during training to minimize the cost function using gradient descent and backpropagation.

It is worth deciding on the neural network training algorithm:

1) Import libraries
2) Data checking
   a) Checking for Lost Values
   b) Data normalization
3) Modeling
4) Getting a result

The neural network analyzes the biometric data by learning patterns and features from the input data during the training process. It can then use these learned features to predict new biometrics. For example, a facial recognition neural network can learn to recognize certain facial features and patterns, such as eye position or mouth shape, and use that information to identify people in new images. Similarly, a fingerprint recognition neural network can learn to recognize the unique ridges and patterns on a fingerprint and use that information to verify a person's identity.

Let's analyze this algorithm on a specific example (Fig. 1).

It is necessary to create and train a neural network based on the MNIST database that will recognize handwritten numbers.

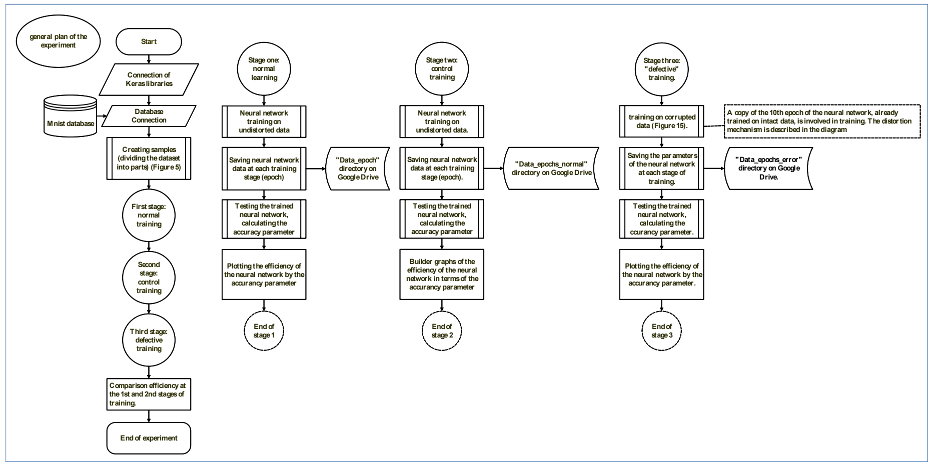

general plan of the experiment
Start
Connection of Keras libraries
Mnist database
Database Connection
Creating samples (dividing the dataset into parts) (Figure 5)
First stage: normal training
Second stage: control training
Third stage: defective training
Comparison efficiency at the 1st and 2nd stages of training.
End of experiment
Stage one: normal learning
Neural network training on undistorted data
Saving neural network data at each training stage (epoch)
"Data_epoch" directory on Google Drive
Testing the trained neural network, calculating the accuracy parameter
Plotting the efficiency of the neural network by the accurancy parameter
End of stage 1
Stage two: control training
Neural network training on undistorted data.
Saving neural network data at each training stage (epoch).
"Data_epochs_normal" directory on Google Drive
Testing the trained neural network, calculating the accuracy parameter
Builder graphs of the efficiency of the neural network in terms of the accurancy parameter
End of stage 2
Stage three: "defective" training.
training on corrupted data (Figure 15).
A copy of the 10th epoch of the neural network, already trained on intact data, is involved in training. The distortion mechanism is described in the diagram
Saving the parameters of the neural network at each stage of training.
"Data_epochs_error" directory on Google Drive.
Testing the trained neural network, calculating the ccurancy parameter.
Plotting the efficiency of the neural network by the accurancy parameter.
End of stage 3

Fig 1. General structure of the experiment algorithm

It is necessary to create and train a neural network based on the MNIST database that will distinguish between handwritten numbers.

The following libraries and modules have been used:

Numpy is a library for working with multidimensional arrays and matrices [23].

Keras is a framework for building and training neural networks. It is part of the Tensorflow library and allows you to create neural network models using a set of high-level abstractions.

Import the NumPy library and give it the alias “np". We import the MNIST dataset from the Keras library. MNIST contains images of numbers from 0 to 9, which will be used to train the neural network [25].

Sequential is a neural network model in which layers are added sequentially one after the other. Dense and Flatten layers from the Keras library. Adam is an optimizer from the Keras library. Optimizers are used to adjust the neural network weights during training. Sparse Categorical Crossentropy is a loss function from the Keras library. The loss function is a metric that evaluates how well a neural network performs on a classification task.

Matplotlib is used for data visualization. ModelCheckpoint is a class that allows you to save model weights during training. The load_model function from the Keras library is used to load a saved model from a file [25].

Next, we connect Google Drive to Google Colab. Google Drive is used to save the model and other files.

The shuttle module in Python provides a high-level interface for working with files and directories. It contains functions for copying, moving, renaming, and deleting files and directories.

```
import numpy as np
from tensorflow.keras.datasets import mnist
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense, Flatten
from tensorflow.keras.optimizers import Adam
from tensorflow.keras.losses import SparseCategoricalCrossentropy
import matplotlib.pyplot as plt
from tensorflow.keras.callbacks import ModelCheckpoint
from keras.models import load_model
from google.colab import drive
drive.mount('/content/gdrive')
import shutil
```

We load the MNIST dataset (Fig. 2) and split it into training and test data, where `x_train` and `x_test` – numpy arrays containing images of handwritten digits (Fig. 3). Each image is a 28x28 matrix of pixels. Each pixel corresponds to a value from 0 to 255 (various shades of black, white and gray).

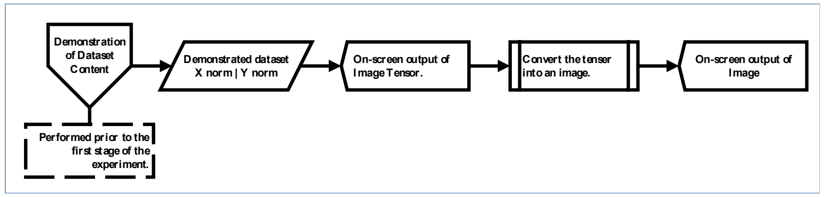

Demonstration of Dataset Content
Demonstrated dataset X norm | Y norm
On-screen output of Image Tensor.
Convert the tenser into an image.
On-screen output of Image
Performed prior to the first stage of the experiment.

Fig 2. The structure of the algorithm of the module for demonstrating the contents of the Mnist dataset

`y_train` and `y_test` – numpy arrays containing labels for the corresponding images in the training and test sets. The label is an integer from 0 to 9 that corresponds to the handwritten digit on the corresponding image.

```
(x_train, y_train), (x_test, y_test) = mnist.load_data()
```

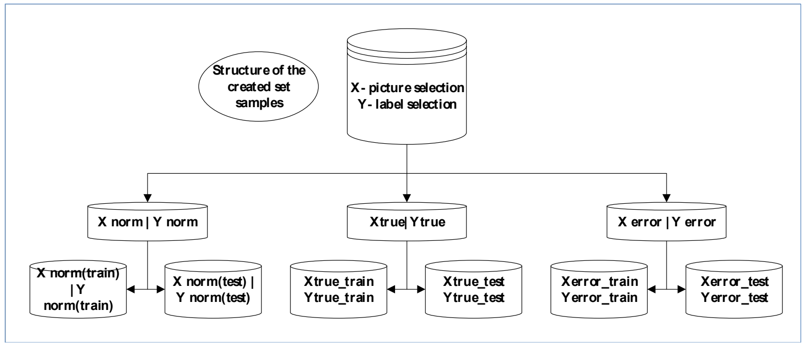
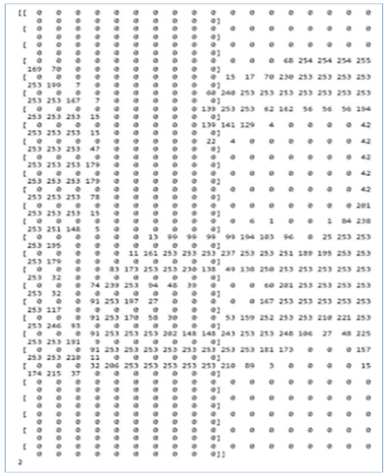


Fig 3. Scheme of splitting the dataset into samples



```
[[  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0   0   0  68 254 254 254 255
  169  70   0   0   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0  15  17  70 230 253 253 253 253
  253 199   7   0   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0  60 240 253 253 253 253 253 253 253
  253 253 167   7   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0 139 253 253  62 162  56  56  56 194
  253 253 253  15   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0 139 141 129   4   0   0   0   0  42
  253 253 253  15   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0  22   4   0   0   0   0   0   0  42
  253 253 253  47   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0  42
  253 253 253 179   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0  42
  253 253 253 179   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0  42
  253 253 253  78   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0 201
  253 253 253  15   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0   6   1   0   0   1  84 238
  253 251 148   5   0   0   0   0   0   0]
 [  0   0   0   0   0   0  13  99  99  99  99 194 103  96   0  25 253 253
  253 195   0   0   0   0   0   0   0   0]
 [  0   0   0   0   0  11 161 253 253 253 237 253 253 251 189 195 253 253
  253 179   0   0   0   0   0   0   0   0]
 [  0   0   0   0  83 173 253 253 230 138  49 138 250 253 253 253 253 253
  253  32   0   0   0   0   0   0   0   0]
 [  0   0   0  74 239 253  94  48  39   0   0   0  60 201 253 253 253 253
  253  32   0   0   0   0   0   0   0   0]
 [  0   0   0  91 253 197  27   0   0   0   0   0 167 253 253 253 253 253
  253 117   0   0   0   0   0   0   0   0]
 [  0   0   0  91 253 170  58  30   0   0  53 159 252 253 253 210 221 253
  253 246  93   0   0   0   0   0   0   0]
 [  0   0   0  91 253 253 253 202 148 148 243 253 253 248 106  27  48 225
  253 253 191   9   0   0   0   0   0   0]
 [  0   0   0  91 253 253 253 253 253 253 253 253 181 173   0   0   0 157
  253 253 210  11   0   0   0   0   0   0]
 [  0   0   0  32 206 253 253 253 253 253 210  89   3   0   0   0   0  15
  174 215  37   0   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0]]
2
```

Fig 4. Output of element 277

For clarity, we derive one of the elements `x_train and y_train` in the form in which they are stored in tuples. Take element number 277 (Fig. 5).

Display the image of the element `x_train[277]` in black and white (white is 0 and black is 255) (Fig. 4).

Fig 5. Image of element 277

Let's normalize the pixel values to bring the values from 0 to 255 to the range from 0 to 1. This is done by dividing each pixel value by 255. Pixel normalization improves the performance of the model, as it facilitates training and reduces the time required for processing data.

At the next stage, we split the training and test datasets into two equal parts. The first part will be used for the first training of the neural network, and the second part will be distorted and used in the second training. Define the architecture of the model (See Fig. 6).

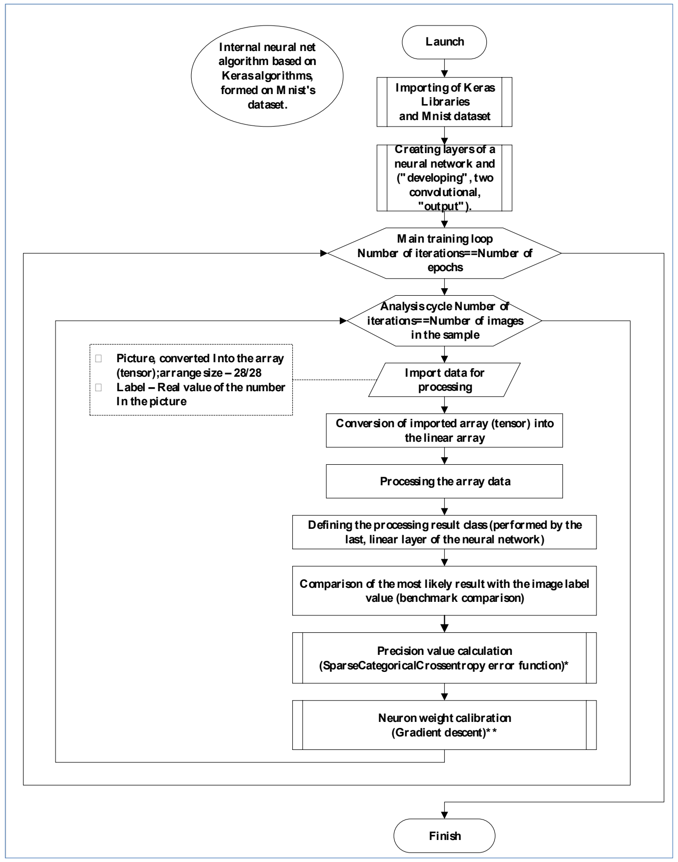

Internal neural net algorithm based on Keras algorithms, formed on Mnist's dataset.
Launch
Importing of Keras Libraries and Mnist dataset
Creating layers of a neural network and ("developing", two convolutional, "output").
Main training loop Number of iterations==Number of epochs
Analysis cycle Number of iterations==Number of images in the sample
Picture, converted Into the array (tensor);arrange size – 28/28
Label – Real value of the number In the picture
Import data for processing
Conversion of imported array (tensor) into the linear array
Processing the array data
Defining the processing result class (performed by the last, linear layer of the neural network)
Comparison of the most likely result with the image label value (benchmark comparison)
Precision value calculation (SparseCategoricalCrossentropy error function)*
Neuron weight calibration (Gradient descent)**
Finish

Fig 6. The structure of the algorithm of the internal mechanism of the neural network

Using the method of experiment, the optimal structure of the neural network and the parameters of its training were selected.

The first layer – Flatten converts a two-dimensional array (28, 28) into a one-dimensional array (dimension 784) so that it can be fed to the input of the neural network.

The second layer – Dense with 100 neurons and the ReLU (Rectified Linear Unit) activation function uses a linear operation followed by a non-linear activation function. The ReLU activation function returns 0 for negative values and the value itself for positive ones.

The third layer is Dense with 50 neurons and the ReLU activation function.

The first optimizer parameter defines the optimization method that will be used to train the model. In this case, the optimizer `Adam` is used at the rate of learning (`learning_rate`) equal to 0.01.

Second parameter `loss` defines the loss function to be used during model training. Here we use categorical cross entropy (`SparseCategoricalCrossentropy`).

The fourth layer is Dense with 10 neurons and softmax activation function. softmax converts neuron values into probabilities summing up to 1.0 and is used for multi-class classification [29].

We compile the model with the necessary parameters.

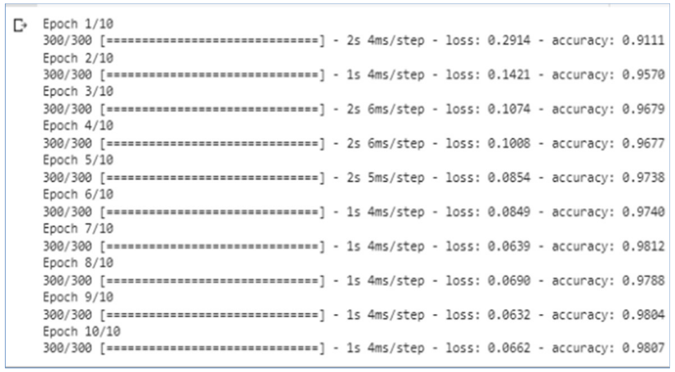
Third parameter `metrics` defines the metrics that will be used to evaluate the model. In this case, we will use only the accuracy metric (`accuracy`).

```
model.compile(optimizer=Adam(learning_rate=0.01),
              loss=SparseCategoricalCrossentropy(),
              metrics=['accuracy'])
```

Create a checkpoint object that is used to save the state of the model as a file in a folder `data_epochs` after each learning epoch.

```
checkpoint = ModelCheckpoint('/content/gdrive/My Drive/neyro/data_epochs/
epoch_{epoch:02d}.h5')
```

Train the model on training data `Xtrue_train` with appropriate labels `Ytrue_train`. Specify the number of epochs 10. For one training iteration, we take the batch size 100 (Fig. 7).



```
Epoch 1/10
300/300 [==============================] - 2s 4ms/step - loss: 0.2914 - accuracy: 0.9111
Epoch 2/10
300/300 [==============================] - 1s 4ms/step - loss: 0.1421 - accuracy: 0.9570
Epoch 3/10
300/300 [==============================] - 2s 6ms/step - loss: 0.1074 - accuracy: 0.9679
Epoch 4/10
300/300 [==============================] - 2s 6ms/step - loss: 0.1008 - accuracy: 0.9677
Epoch 5/10
300/300 [==============================] - 2s 5ms/step - loss: 0.0854 - accuracy: 0.9738
Epoch 6/10
300/300 [==============================] - 1s 4ms/step - loss: 0.0849 - accuracy: 0.9740
Epoch 7/10
300/300 [==============================] - 1s 4ms/step - loss: 0.0639 - accuracy: 0.9812
Epoch 8/10
300/300 [==============================] - 1s 4ms/step - loss: 0.0690 - accuracy: 0.9788
Epoch 9/10
300/300 [==============================] - 1s 4ms/step - loss: 0.0632 - accuracy: 0.9804
Epoch 10/10
300/300 [==============================] - 1s 4ms/step - loss: 0.0662 - accuracy: 0.9807
```

Fig 7. Data output after the first training

`shuffle=True` specifies that the training dataset will be shuffled before each epoch to avoid the possibility that the model might remember the order of the training examples.

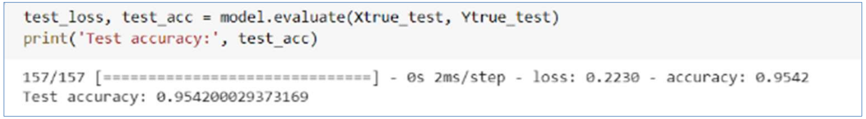
`callbacks=[checkpoint]` indicates that the object `checkpoint` will be used as a callback to save the state of the model after each epoch.

Learning outcomes are saved to an object `history`. After training the model, history will contain information about the change in the loss function and accuracy metrics during model training.

```
epochs = 10
```

```
history = model.fit(Xtrue_train, Ytrue_train, epochs=epochs,
                    batch_size=100, shuffle=True, callbacks=[checkpoint])
```

Let's evaluate the model on test data. After the first training, we got an accuracy of 0.9542 on test data.



```
test_loss, test_acc = model.evaluate(Xtrue_test, Ytrue_test)
print('Test accuracy:', test_acc)

157/157 [==============================] - 0s 2ms/step - loss: 0.2230 - accuracy: 0.9542
Test accuracy: 0.954200029373169
```

Fig 8. Accuracy of training on test data

Derive the graphs of the first training.

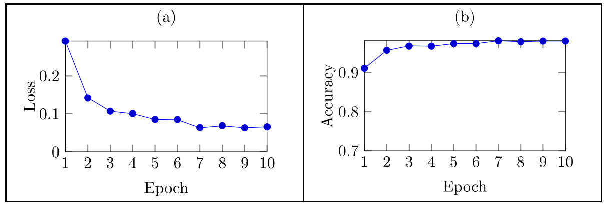


Fig 9. a) Graph of losses, b) Graph of accuracy

Thus, the neural network was trained to recognize handwritten numbers on the MNIST dataset.

The next stage of work consists in distorting the dataset, that is, simulating the intervention of an attacker, and then analyzing the results after retraining the neural network.

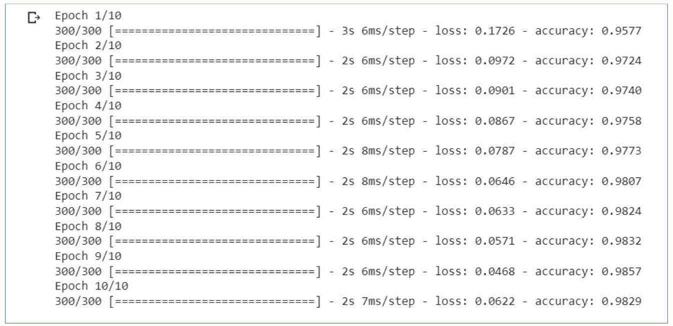
The task is to find a way to calculate malicious interference in the operation of the neural network.

Let's put forward a hypothesis that one of the signs of interference in the neural network may be the difference in performance and accuracy to the trained neural network from the original one.

Let us test this hypothesis on a practical problem.

Before training the model on a distorted dataset, we will make a copy of the 10th epoch file and train the model loaded from it in the same way, but on the second half of the dataset (`Xerror_train` и `Yerror_train`), but without changing anything in it.

```
source_file = '/content/gdrive/My Drive/neyro/data_epochs/epoch_10.h5'
destination_file = '/content/gdrive/My Drive/neyro/data_epochs/
                   epoch_10_copy1.h5'
      # Copying a file to a new path
shutil.copyfile(source_file, destination_file)
model = load_model('/content/gdrive/My Drive/neyro/data_epochs/
                   epoch_10_copy1.h5')
      #Model         compilation         with         Adam         optimizer,
        SparseCategoricalCrossentropy loss function and accuracy metric.
history = model.compile(optimizer=Adam(learning_rate=0.01),
              loss=SparseCategoricalCrossentropy(),
              metrics=['accuracy'])
      # Creating a folder to save training data after each epoch
import os
if not os.path.exists('/content/gdrive/My Drive/neyro/
                      data_epochs_continue'):
    os.makedirs('/content/gdrive/My Drive/neyro/data_epochs_continue')
      # Saving the model after each epoch to folder data_epochs_error
checkpoint = ModelCheckpoint('/content/gdrive/My Drive/
                      neyro/data_epochs_continue/epoch_{epoch:02d}.h5')
      # Training the model on the second half (not yet distorted) of the
dataset
epochs = 10
history = model.fit(Xerror_train, Yerror_train, epochs=epochs,
                    batch_size=100, shuffle=True, callbacks=[checkpoint])
```



```
Epoch 1/10
300/300 [==============================] - 3s 6ms/step - loss: 0.1726 - accuracy: 0.9577
Epoch 2/10
300/300 [==============================] - 2s 6ms/step - loss: 0.0972 - accuracy: 0.9724
Epoch 3/10
300/300 [==============================] - 2s 6ms/step - loss: 0.0901 - accuracy: 0.9740
Epoch 4/10
300/300 [==============================] - 2s 6ms/step - loss: 0.0867 - accuracy: 0.9758
Epoch 5/10
300/300 [==============================] - 2s 8ms/step - loss: 0.0787 - accuracy: 0.9773
Epoch 6/10
300/300 [==============================] - 2s 8ms/step - loss: 0.0646 - accuracy: 0.9807
Epoch 7/10
300/300 [==============================] - 2s 6ms/step - loss: 0.0633 - accuracy: 0.9824
Epoch 8/10
300/300 [==============================] - 2s 6ms/step - loss: 0.0571 - accuracy: 0.9832
Epoch 9/10
300/300 [==============================] - 2s 6ms/step - loss: 0.0468 - accuracy: 0.9857
Epoch 10/10
300/300 [==============================] - 2s 7ms/step - loss: 0.0622 - accuracy: 0.9829
```

Fig 10. Output of training results

Accuracy after retraining increased from 0.9542 to 0.9762 (Fig. 10).

Let's create a copy of the epoch 10 file so that it can be used in the second training of the neural network and compare the received data.

Let's compile the model using the same parameters as the first time.

```
history = model.compile(optimizer=Adam(learning_rate=0.01),
                loss=SparseCategoricalCrossentropy(),
                metrics=['accuracy'])
```

Training and testing of the ANN is carried out on the second half of the MNIST dataset (with the prefix error), which was created at the beginning of the code.

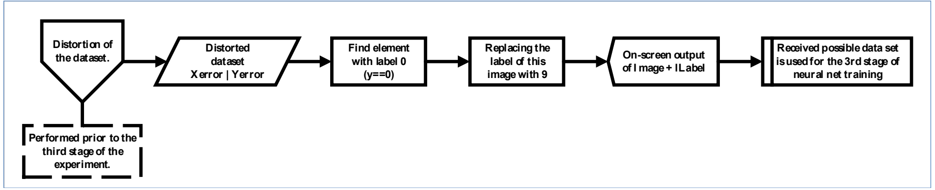
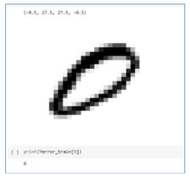

Distortion of the dataset.
Distorted dataset Xerror | Yerror
Find element with label 0 (y==0)
Replacing the label of this image with 9
On-screen output of Image + ILabel
Received possible data set is used for the 3rd stage of neural net training
Performed prior to the third stage of the experiment.

Fig 11. The structure of the algorithm for direct distortion of the sample "Error"

Let's find an element from the data set whose value is 0 and display its image and its corresponding label.`plt.imshow(Xerror_train[9], cmap='binary')` (Fig. 12).



```
(-0.5, 27.5, 27.5, -0.5)
```

```
[ ] print(Yerror_train[9])
    0
```

Fig 12. Image of dataset element 0 and its label

Replace all label values in the training dataset `Yerror_train` with values 0 to 9 (Fig. 11).

This stage is an imitation of an attacker's action. This example can be projected onto real face recognition neural networks, that is, different dataset numbers are different faces of people.

One of the reasons for the emergence of the Adversarial attack is that machine learning methods were originally developed for stationary and secure environments, where the training and test samples were generated from the same statistical distribution. However, in practice, attackers can covertly manipulate input data to exploit vulnerabilities in ML algorithms and compromise the security of the entire machine learning system.

The type of attack in which an attacker seeks to gain access to the data and the learning process of an ML model to "poison" it (train it incorrectly) for subsequent inadequate work is called poisoning. It can be seen as a malicious infection of the training data. Thus, the "white box" strategy is used here, when the attacker has information about the victim – "malicious knowledge" (Adversarial Knowledge, AK): how the data for training are prepared and from what sources and what they are, what are the main functions of the attacked system, what algorithms it uses, what are the results, etc. Poison attacks involve insider information about the ML system and a high level of attacker's competence in Data Science [23].

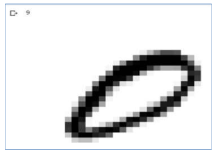
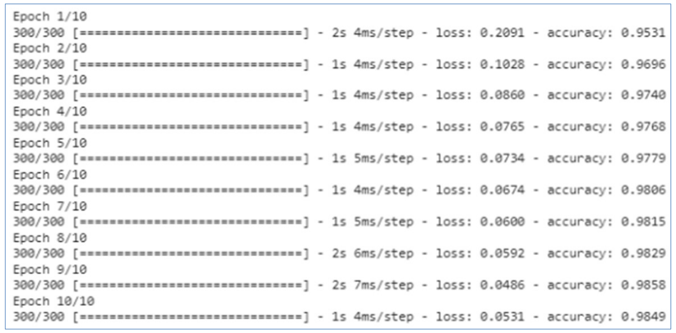


Fig 13. Image of dataset element 0 and its label after replacement

As a result, all images 0 correspond to labels with a value of 9 (Fig. 13).

Retrained the neural network on data sets `Xerror_train` and `Yerror_train`, using the same parameters as the first time.

```
epochs = 10
history = model.fit(Xerror_train, Yerror_train, epochs=epochs,
                    batch_size=100, shuffle=True, callbacks=[checkpoint])
```



```
Epoch 1/10
300/300 [==============================] - 2s 4ms/step - loss: 0.2091 - accuracy: 0.9531
Epoch 2/10
300/300 [==============================] - 1s 4ms/step - loss: 0.1028 - accuracy: 0.9696
Epoch 3/10
300/300 [==============================] - 1s 4ms/step - loss: 0.0860 - accuracy: 0.9740
Epoch 4/10
300/300 [==============================] - 1s 4ms/step - loss: 0.0765 - accuracy: 0.9768
Epoch 5/10
300/300 [==============================] - 1s 5ms/step - loss: 0.0734 - accuracy: 0.9779
Epoch 6/10
300/300 [==============================] - 1s 4ms/step - loss: 0.0674 - accuracy: 0.9806
Epoch 7/10
300/300 [==============================] - 1s 5ms/step - loss: 0.0600 - accuracy: 0.9815
Epoch 8/10
300/300 [==============================] - 2s 6ms/step - loss: 0.0592 - accuracy: 0.9829
Epoch 9/10
300/300 [==============================] - 2s 7ms/step - loss: 0.0486 - accuracy: 0.9858
Epoch 10/10
300/300 [==============================] - 1s 4ms/step - loss: 0.0531 - accuracy: 0.9849
```

Fig 14. Data output after training

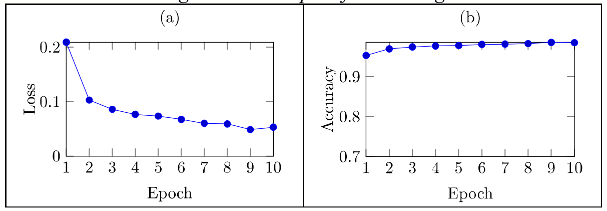


Fig 15. a) Graph of losses b) Graph of accuracy

We got an accuracy on the test data of 0.8758, which is much less than the result of 0.9762 obtained after the model was trained on an undistorted dataset, and less than after the first training, where the result on the test data was 0.9542 (Fig. 14, 15, Table 1).

Table. 1. Simulation results

	ANN training		
	First	Undistorted dataset	Distorted dataset
Accuracy	0,9542	0,9762	0,8758

This confirms the hypothesis. Thus, when retraining a neural network on a distorted dataset, the accuracy on test data drops, and when retraining on an undistorted dataset, it increases.

This is since the information received by the neural network about handwritten numbers from the MNIST dataset during initial training comes into conflict with the information received during re-training on the dataset with changed labels.

Therefore, a sharp decrease in the accuracy of the neural network during additional training is one of the signs of an attacker's intervention.

The presented method is one of the simplest methods for detecting malicious interference, and in further research it is planned to rely on the gradient descent method, which can be more accurate for this task. This is exactly what scientists from Cornell University did for the article “Interpreting Deep Neural Networks with SVCCA” [30].

4. Conclusion

In conclusion, adversarial attacks are of particular interest in the field of artificial intelligence, and a large amount of research has been done in this direction. While defense strategies have been proposed and the arms race between attackers and defenders continues, new methods of attack are constantly being developed. Thus, it is important for researchers to study adversarial attacks and develop new protection strategies to ensure the security of machine learning systems, the formation of a secure model of trusted artificial intelligence.

References / Список литературы

[1] Marshalko G. Attacks on biometric systems. Information Security (in Russian) / Маршалко Г. Атаки на биометрические системы. Information Security. Available at: https://www.itsec.ru/articles/ataka-na-biometricheskie-sistemy, accessed 30.03,2023.

[2] Ivanyuk V. Neural Networks and Their Analysis. Chronoeconomics, 2021, issue 4, pp. 58-61 / Иванюк В.А. Нейронные сети и их анализ. Хроноэкономика, вып. 4, 2021 г., стр. 58-61.

[3] Kachagina K. S., Safarova A. D. Neron Networks - Development Prospects. E-Scio, issue 2, 2021, 10 p. (in Russian) / Качагина К.С., Сафарова А.Д. Нейронные сети - перспективы развития. E-Scio, вып. 2, 2021 г., 10 стр.

[4] Akhtar Z., Foresti G. L. Face spoof attack recognition using discriminative image patches. Journal of Electrical and Computer Engineering, 2016, article id 4721849, 15 p.

[5] Chernobrov A. How to cheat a neural network or what is an Adversarial attack (in Russian) / Чернобров А. Как обмануть нейросеть или что такое Adversarial attack. Available at: https://chernobrovov.ru/articles/kak-obmanut-nejroset-ili-chto-takoe-adversarial-attack.html, accessed: 02.04.2023.

[6] Goodfellow I.J., Shlens J., Szegedy C. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014, 11 p.

[7] Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial examples in the physical world. arXiv preprint arXiv:1607. 2016, 14 p.

[8] Wiyatno R., Xu A. Maximal Jacobian-based Saliency Map Attack. arXiv preprint arXiv:1808.07945, 2018, 5 p

[9] Moosavi-Dezfool S.-M., Fawzi A., Frossard P. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2574-2582

[10] Madry A., Makelov A. et al. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083, 2017, 28 p.

[11] Papernot N., McDaniel P. et al. The limitations of deep learning in adversarial settings. In Proc. of the IEEE European Symposium on Security and Privacy (EuroS&P), 2016, pp. 372-387.

[12] Carlini N., Wagner D. Adversarial examples are not easily detected: Bypassing ten detection methods. In Proc. of the 10th ACM Workshop on Artificial Intelligence and Security, 2017, pp. 3-14.

[13] Akhtar N., Mian A. Threat of adversarial attacks on deep learning in computer vision: A survey. IEEE Access, vol. 6, 2018, pp. 14410-14430.

[14] Xiao C., Deng R. et al. Characterizing adversarial examples based on spatial consistency information for semantic segmentation. Lecture Notes in Computer Science, vol. 11214, 2018, pp. 220-237.

[15] Xie C., Wang J. et al. Adversarial examples for semantic segmentation and object detection. In Proc. of the IEEE International Conference on Computer Vision, 2017, pp. 1369-1378.

[16] Samangouei P., Kabkab M., Chellappa R. Defense-gan: protecting classifiers against adversarial attacks using generative models. arXiv preprint arXiv:1795.06605, 2018, 18 p.

[17] Xu H., Ma Y. et al. Adversarial Attacks and Defenses in Images, Graphs and Text: A Review. International Journal of Automation and Computing, vol. 17, issue 2, 2020, pp. 151-178.

[18] Xu W., Evans D., Qi Y. Feature squeezing: Detecting adversarial examples in deep neural networks. In Proc. of the Network and Distributed Systems Security (NDSS) Symposium, 2018, 15 p

[19] Wang Y., Kang L. et al. Fingerprint presentation attack detection using convolutional neural network with transfer learning. IEEE Access, vol. 7, 2019, pp. 131443-131451.

[20] Nguyen T.M., Kim K.H. et al. Generative adversarial network-based face presentation attack detection using partial convolution and multi-domain learning. IEEE Transactions on Information Forensics and Security, vol. 14, issue 10, 2019, pp. 2764-2779.

[21] Li X., Chen T., Yang J.. Adversarial fingerprint attacks and defenses. IEEE Transactions on Information Forensics and Security, vol. 14, issue 1, 2019, pp. 66-80.

[22] Tan H., Li H., et al. Deep learning based liveness detection: A survey. ACM Computing Surveys, vol. 52, issue 3, (2019), pp. 1-27.

[23] The official website of the NumPy Library. Available at: Available: https://numpy.org, accessed 30.03.2023.

[24] Examples of neural networks' implementation (in Russian) / Примеры реализации нейронных сетей. Available at: Available: https://webtort.ru, accessed 30.03.2023.

Information about authors / Информация об авторах

Denis Vladimirovich KOTLYAROV is a student. His research interests include artificial neural networks, cryptography.

Денис Владимирович КОТЛЯРОВ – студент. Его научные интересы включают искусственные нейронные сети, криптография.

Gleb Dmitrievich DYUDYUN – student. His research interests include artificial neural networks, cryptography.

Глеб Дмитриевич ДЮДЮН – студент. Его научные интересы включают искусственные нейронные сети, криптография.

Natalya Vitalievna RZHEVSKAYA is a student. Her research interests include artificial neural networks, cryptography.

Наталья Витальевна РЖЕВСКАЯ – студентка. Ее научные интересы включают искусственные нейронные сети, криптография.

Maria Anatolyevna LAPINA – Candidate of Physical and Mathematical Sciences, Associate Professor of the Department of Information Security of Automated Systems. Her research interests include artificial neural networks, cryptography, numerical methods, differential calculus.

Мария Анатольевна ЛАПИНА – кандидат физико-математических наук, доцент кафедры информационной безопасности автоматизированных систем. Ее научные интересы включают искусственные нейронные сети, криптография, численные методы, дифференциальные исчисления.

Mikhail BABENKO – Doctor of Physical and Mathematical Sciences, Head of the Department of Computational Mathematics and Cybernetics. His research interests include artificial neural networks, cryptography, homomorphic encryption, secret sharing schemes, modular arithmetic, cloud computing.

Михаил Григорьевич БАБЕНКО – доктор физико-математических наук, заведующий кафедрой вычислительной математики и кибернетики. Его научные интересы включают искусственные нейронные сети, криптография, гомоморфное шифрования, схемы разделения секрета, модулярная арифметика, облачные вычисления.