e-ISSN 1607-3274 Радіоелектроніка, інформатика, управління. 2019. № 2
p-ISSN 2313-688X Radio Electronics, Computer Science, Control. 2019. № 2

UDC 004.93

SCOPING ADVERSARIAL ATTACK FOR IMPROVING ITS QUALITY

Khabarlak K. S. – Student of the Department of System Analysis and Control, Dnipro University of Technology, Dnipro, Ukraine.

Koriashkina L. S. – PhD, Associate Professor of the Department of System Analysis and Control, Dnipro University of Technology, Dnipro, Ukraine.

ABSTRACT

Context. The subject of this paper is adversarial attacks, their types, reasons for the emergence. A simplified fast and effective logistic regression attack algorithm has been presented. The work's relevance is explained by the fact that neural network's critical vulnerability the so-called adversarial examples is yet to be deeply explored. By exploiting such a mechanism, it is possible to get a deliberate result from it breaking defenses of neural-network-based safety systems.

Objective. The purpose of the work is to develop algorithms for different kinds of attacks of a trained neural network with respect to preliminary the network's weights analysis, to estimate attacked image quality loss, to perform a comparison of the developed algorithms and other adversarial attacks of a similar type.

Method. A fast and fairly efficient attack algorithm that can use either whole image or its certain regions is presented. Using the SSIM image structural similarity metric, an analysis of the algorithm and its modifications was carried out, as well as a comparison with previous methods using gradient for the attack.

Results. Simplified targeted and non-targeted attack algorithms have been built for a single-layer neural network trained to perform handwritten digit classification on the MNIST dataset. A visual and semantic interpretation of weights as pixel "importance" for recognizing an image as one class or another is given. Based on structural image similarity index SSIM an image quality loss analysis has been performed for images attacked by the proposed algorithms on the whole test dataset. Such an analysis has revealed the classes the most vulnerable to an adversarial attack as well as images, whose class can be changed by adding noise imperceptible by a human being.

Adversarial examples built with the developed algorithm has been transferred to a 5-layered network of an unknown architecture. In many cases images that were difficult to attack for the original network have seen a higher transfer rates, then the ones needed only minor image changes.

Conclusions. Adversarial examples built upon the adversarial attack scoping idea and the methodic of the input data analysis can be easily generalized to other image recognition problems which makes it applicable to a wide range of practical tasks. This way, another way of analyzing neural network safety (logistic regression included) against input data attacks is presented.

KEYWORDS: adversarial attacks, fast adversarial attack algorithm, logistic regression, neural network vulnerabilities.

ABBREVIATIONS

FGSM Fast Gradient Sign Method;

L-BFGS Limited-memory BFGS (Broyden-Fletcher-Goldfarb-Shanno) algorithm;

MAE Mean Absolute Error;

MNIST Modified National Institute of Standards and Technology;

PSNR Peak Signal-to-Noise Ratio;

RP2 Robust Physical Perturbations;

Softmax Softened (via exponent) max function.

SSIM Structural Similarity Index;

NOMENCLATURE

b – biases vector;

I – feature space dimensionality, neuron count in the network's input layer;

image – source image;

K – number of classes, neuron count in the output layer;

M – training batch size, used to define number of images used in a single optimization method step during training;

max_*steps* – number of attack steps;

min_*difference* – minimal difference between classes' weights allowed for a pixel attack;

S – source class;

source_weights – trained network's weight matrix for a class predicted by the neural network for the source image;

step – value, which defines pixel brightness change on a given iteration;

T – target class (desired result);

target_weights – trained network weight matrix for the target classification class;

W – network's weight matrix;

x – network's input vector, image pixel brightness vector;

y – unit vector, which defines object attribution to one class or another;

z – network's output vector;

$\hat{z}$ – desired (target) neural network's output;

α – algorithm parameter;

$\sigma_k(\cdot)$ – output layer's k^{th} neuron softmax activation function, which computes layer output by its input.

INTRODUCTION

An increasing number of tasks is being solved with neural-network-based solutions. Neural networks have reached the dominant position in image recognition since 2012, when in ImageNet Large Scale Visual Recognition Challenge AlexNet has got first place with a large margin [1]. A growing need of analyzing neural networks and their vulnerabilities arises with a more prominent use in security, video surveillance systems, self-driving cars and robots.

Pretrained neural networks are being used by many companies to reach their goals. A set of security vulnerabilities is disclosed due to a wide spread of neural networks with similar architecture that are trained on

DOI 10.15588/1607-3274-2019-2-12

publicly available datasets. Given a slight modification of an input data small enough to be imperceptible by a human being it is possible to make the neural network misclassify the data or even output some specific class. These are known as non-targeted and targeted adversarial attacks respectively.

By exploring trained neural network attack algorithms, understanding attack preconditions and structural image changes caused by it, it will be possible to make neural networks more efficient and robust to the input data perturbations.

In most cases the process of neural network inference can be treated as a black box even though its architecture and training dataset is known, giving a clear interpretation of underlying weights is a difficult problem. Lately a range of research papers with an attempt to track and formalize network inference model through its input data and with the use of optimization and statistics theories has been published (i.e. [2–4]). Despite that fact, development of the toolchain for understanding and diagnosing machine learning is still a prioritized research problem.

The object of study is the neural network attack process.

The subject of study is the types of adversarial attacks, consequences caused by them and possible reasons of their existence.

The purpose of the work is to: 1) develop different attack algorithms on a single-layer trained neural network given the results of a preliminary analysis of the network's weights; 2) estimation of image quality loss after the modification; 3) comparison of the attack results conducted by the developed algorithms with different adversarial attack types that were previously published in the research papers.

1 PROBLEM STATEMENT

Given a training set

$$X = \left\{ x^{(j)}, y^{(j)} \right\}_{j=1}^{J} : \; x^{(j)} \in R^I; \;\; y^{(j)} \in R^K,$$

$$y_k^{(j)} = 0 \vee 1 \; \forall j, \sum_{k=1}^{K} y_k^{(j)} = 1 \; \forall j .$$

1. Find a weight matrix $W : [I \times K]$ and a vector $b = (b_1, b_2, ..., b_K)$, which minimize function

$$G(W,b) = \frac{1}{M} \sum_{m=1}^{M} \gamma(W,b; z^{(m)}, y^{(m)}) =$$

$$= -\frac{1}{M} \sum_{m=1}^{M} \sum_{k=1}^{K} y_k^{(m)} \log\left(\sigma_k \left(x^{(m)} \cdot W + b \right) \right),$$

where

$$\gamma(W,b; z, y) = -\sum_{k=1}^{K} y_k \log(z_k),$$

$$z = \sigma(x \cdot W + b),$$

$$\sigma(z) = (\sigma_1(z), \sigma_2(z), ..., \sigma_K(z)), \; \sigma_k(z) = e^{z_k} \Big/ \sum_{j=1}^{K} e^{z_j} .$$

2. Find such a perturbation $\Delta x \in R^I$, which for a given value $T \in \{1, 2, ..., K\}$ and deliberately chosen x, such that $z_S = \max\limits_{k=1,K} z_k, S \neq T$,

$$\hat{z}_T = \max_{k=1,K} \hat{z}_k ,$$

where $\hat{z} = \sigma((x + \Delta x) \cdot W + b)$.

3. Find such a perturbation $\Delta x \in R^I$, which for a deliberately chosen x, such that $z_S = \max\limits_{k=1,K} z_k$, $l \in \{1, 2, ..., K\}$ can be found that meets a condition of $\hat{z}_l > \hat{z}_S$.

2 REVIEW OF THE LITERATURE

The term "Adversarial attack" has been first introduced in the work [3], where it has been shown how with a minor change of an image's pixels leads to a misclassification during neural network inference procedure. What makes the problem more severe is that in many cases these changes cannot be observed by a human. As it has been shown by the authors, such a perturbation is not a random network's training issue. The very same modified image can be transferred with a success to a different network trained on a different but a similar dataset. To archive that goal a Box-constrained L-BFGS algorithm has been introduced. Another algorithm FGSM (Fast Gradient Step Method) has been developed as an enhancement, where a first-order approximation of the loss function is used to generate adversarial examples [6]. I-FGM is an iterative algorithm that builds on top of the ideas introduced previously and uses gradient of the loss function. These algorithms have a simpler interpretation and are faster to perform the attack. In 2017 an international neural network defense and attack competition has been held by the Google Brain team. The competition's winners have developed an algorithm to attack neural networks with a known architecture. As it has been shown, the generated images have been able not only to mislead the target network, but also other networks trained on other datasets of a similar kind or networks with another architecture. In all the cases attacks were performed on a bitmap to retain the perturbations untouched.

A lot of modern computer vision systems that are being used in security critical domains use deep neural networks behind the scenes. That's why many of the research papers target a scenario of a real-world attack. As such, paper [9] offers a road sign attack algorithm RP2. By placing stickers and drawing graffiti on the real signs, it has been made possible to force the network to misclassify "Stop" sign as a "Speed limitation". Moreover, the perturbations have been shown to be robust to the angle and distance change to a camera.

DOI 10.15588/1607-3274-2019-2-12

Other algorithms of generating reliable physical adversarial perturbations are shown in papers [10, 11 and others].

An efficient black box attack algorithm ZOO has been presented in the paper [12]. This type of an attack can be conducted without any knowledge of the neural network inner workings, the only requirement is to have access to its inference engine: one should be able to send input data and get back class probability distribution. A method of stochastic coordinate descent has been used to optimize the target function. The coordinate to be updated next is chosen by utilizing computed gradient and hessian. Gradient components are calculated with a finite difference method. To further optimize the speed, attack is performed in a hierarchical method, where small scaled images are attacked first and get enlarged over time.

An algorithm to generate efficient targeted adversarial images using optimization methods has been proposed in [13] (C&W attack).

Implementations of the described algorithms as well as some other can be found in an opensource library cleverhans[14].

Different contemporary methods of generating adversarial examples in the field of deep learning were explored and summarized in [15]. Where classification of attacks by characteristics, target goal and features has been presented. A vast review of a research held in the field of machine learning attacks, analysis of the adversarial example precursors and defense methods against them was done in [16].

An attempt to build simplified yet efficient adversarial attack methods on a logistic regression trained on the MNIST dataset has been made in this work. We are exploring two types of white box attacks (based on an assumption of full network weights and architecture knowledge): targeted and non-targeted. Such a choice has been motivated by the following factors:

1) a simplicity of a network's configuration and interpretability of its weights as of an importance of image pixels towards recognizing a picture as a sample of this or that class;

2) an ability of adversarial examples to efficiently highjack neural networks different to the one for which they were generated (as noted in [17]).

3 MATERIALS AND METHODS

The MNIST dataset will be used for the neural network training. Among its advantages is a small size and ability to make accurate predictions even using simple neural networks. The dataset consists of handwritten digits 0–9 of size 28×28. Each digit is a normalized grayscale image. The training subset consists of 60.000 examples, the testing subset 10.000. Both contain samples of digits handwritten by distinct people. Let's unroll images into single-dimensional vector and assume each pixel to be a separate input feature. As a prepossessing step the pixel intensities are normalized into [0, 1] range, which is done by dividing its values by 255 (0 stands for black pixels, 1 for the white ones).

A single-layer neural network is built with an input layer of $I=784$ neurons (by the number of pixels in the unrolled image) and $K=10$ in the output layer (by the number of classes).

Should be noted that during mathematical neural network training problem statement *soft*max activation function was not chosen at random. As is known, *softmax* serves a goal of transforming an arbitrary real-valued vector into a probability distribution of the inferred classed. For example, a network can identify a picture as 8 with a probability 0.9 and as 6 with a likelihood of 0.1. Considering such a "probabilistic" classification during neural network training, cross-entropy loss $\gamma(W,b;z,y)$ has been selected as an error metric between computed outputs y and desired z.

By performing gradient descent for minimizing $G(W,b)$ function, in 12 epochs (roughly 40 seconds worth of training time) an accuracy score of 97.0% and 92.7% on training and test sets has been achieved.

As it has been noted above, the logistic regression's key advantage, which will be used further down to build an attack algorithm, is interpretability of a weight matrix W_{ik} as of an importance or a contribution of i^{th} image pixel towards k^{th} class classification. Precisely, if $W_{ik}>0$, it is expected that an increase of pixel brightness by a some $\delta>0$ will lead to a higher confidence towards classifying an image as an example of k^{th} class, and if the weight $W_{ik}<0$ its decrease will lead to a decay of the probability.

Representation of all inferred classes in a form of a pixel importance map towards classifying each image as an instance of i^{th} class is shown on Fig. 1.

The presented illustrations can be thought of as some generic neural network digit representation.

As shown by the detailed weight matrix analysis: if its element W_{ik} has a large enough positive value (relatively to other matrix elements), then i^{th} image pixel "whiteness" is important for classification of a digit as an instance of k^{th} class; in the opposite case, when W_{ik} is large enough by modulo, but is negative, then black regions are important for k^{th} class. W_{ik} that is close to zero means that color of that pixel has no importance for classification towards k^{th} class. Following the above-described logic, we can use an element-wise multiplication to get pixels equally important for an image classification as of an instance of both classes and element-wise subtraction to get regions whose pixel brightness is more important for one class than the other. Fig. 2 has an example of an element-wise multiplication for digits 0 and 8, where light regions are equally important for both classes.

The result of a subtraction $W_{i8}-W_{i0}, \forall i$ is shown on Fig. 3. It is easily seen that in this case light tones signify pixel importance for classifying 8, the dark ones for classifying 0.

DOI 10.15588/1607-3274-2019-2-12

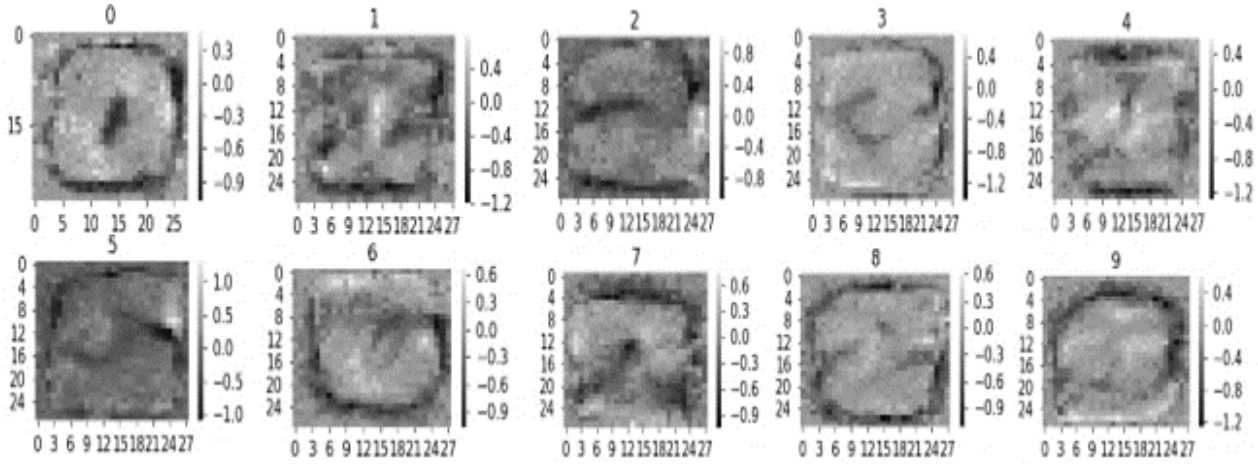

Figure 1 – Image pixel weights for each dataset digit

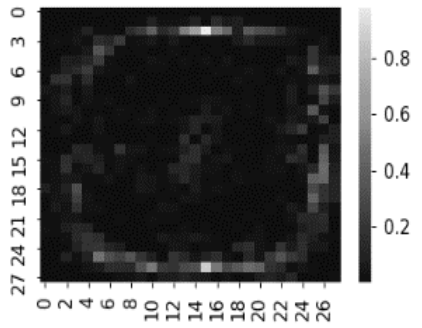

Figure 2 – Element-wise product $W_{i0} \cdot W_{i8}$

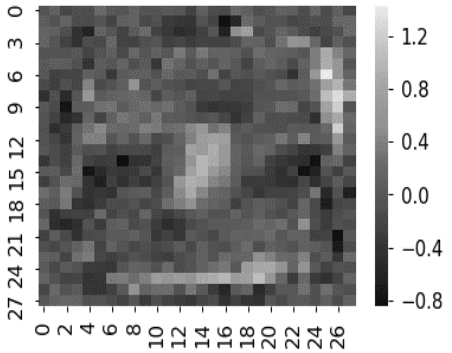

Figure 3 – Pixel importance difference for classes 8 and 0

Similar products and subtractions can be obtained for all pairs of classified examples.

Obviously, subtraction matrices are more significant for performing a targeted attack. By increasing pixel brightness in regions with a large difference (light regions on the figure), one can increase probability of classification of an image as of an instance of the subtrahend and decrease as of an example of the minuend.

To estimate image quality loss L_∞-norm has been used in [8], that is the largest deviation of a pixel brightness over entire image. For images, whose pixel values are bound in the range $[0, 255]$ deviations up to 15 points were permitted. However, such a metric allows to generate nearly unrecognizable (when compared to the source) images, which is not something we are up to. So, a metric that is highly correlated with a human perception is needed. The best results can be obtained by using one of the following metrics: MAE (Mean Absolute Error), PSNR (Peak Signal to Noise Ratio), SSIM (Structural Similarity Index). The first two are easy to compute and are frequent to be used, but they do not take human vision features into account. SSIM metrics has been introduced as an improvement on top of MAE and PSNR and is tuned to match human visual perception system as it is shown in [18]. SSIM metric values lie in range $[-1,1]$. The maximum value signifies that images are identical. This is the metrics that is to be used for an algorithm's quality estimation.

Let's denote neural network attack problem statement. The output is presented as a probability distribution of handwritten digit classes z. Consider that the network predicts image as an instance of class $S \in \{0,1,2,...,9\}$ if $z_S = \max\limits_{k=1,K} z_k$. By changing some pixels' brightness, we want to change neural network prediction to $T \in \{0,1,2,...,9\}, T \neq S$, i.e. $\hat{z}_T = \max\limits_{k=1,K} \hat{z}_k$, $\hat{z} = \sigma((x + \Delta x) \cdot W + b)$. In so doing, we enforce image correctness by clamping brightness values into a range $x_i \in [0,1],\ i = \overline{1,784}$.

Single algorithm step pseudocode:

attack_step (*image*, source_weights, target_weights,
$\alpha \in \{0; 0.5; 1\}$, $min\_difference > 0$,
$step > 0$)

for each point *i* in the *image*
find corresponding weights W_S and W_T in weight matrices for source *S* and target *T* classes
let $delta = \alpha W_T - (1 - \alpha) W_S$
if $|delta| > min\_difference$, then
$x = x + step \cdot delta$
if $x<0$, then $x=0$
if $x>0$, then $x=1$.

Where *image* is an image obtained on the previous algorithm step or the source one if this is the first

DOI 10.15588/1607-3274-2019-2-12

algorithm step; *source_weights* is the trained network weight matrix for original image label; *target_weights* is the trained network weight matrix for a class, towards which we want to change the prediction; *min_difference* is the minimal difference between classes' weight matrices for a pixel to be an attack target; step signifies pixel brightness change on the current iteration; $\alpha \in \{0; 0.5; 1\}$ defines an algorithm modification.

Note, that together with α, *step* and *min_difference*, the number of algorithm steps *max_steps* could also be treated as algorithm parameter. Recommended values for the described parameter values are to follow.

As it was remarked above, by the constraints on the target class j, adversarial attacks are divided into two subtypes:

– if the goal is to assign to an image class j instead of class k, then such an attack is called targeted. This type of attack can be accomplished via the described algorithm with $\alpha = 0.5$ or α=1. With that, it is said that the algorithm has succeeded to attack an image only if the algorithm has been able to change neural network's predicted class into digit j in a finite number of steps, and has failed in all other cases;

– if the goal is to reassign classification of an image of a class k, to any different one $j \neq k$, then the attack is called non-targeted. The algorithm parameter $\alpha = 0$ can be used to perform such an attack. The result is a success if an incorrect classification has been achieved in a finite number of steps.

4 EXPERIMENTS

By using the developed algorithm, an attack $0 \rightarrow 8$ is performed with a goal to force a neural network to misclassify 0 as 8. A generalized pixel importance matrix is used for that (Fig. 3). Adversarial attack results are shown on Fig. 4. Algorithm parameters: $\alpha = 0.5$, *min_difference* = 0.0, *step* = 0.01, *max_steps*=10. As is seen, the attack has succeeded: the digit has been classified as 8 with a probability 44%, in the meantime the source and modified images are virtually the same, which is approved by the SSIM metrics value 0.934.

It should be considered, that if all image pixels are being attacked by bearing in mind only the sign of the attack difference (as it is done in [6]), then the attack will still be successful. However, in such a case image noise can be viewed easily. Moreover, for an attacked image that is classified as 8 with a probability 0.396, SSIM metric value drops down to 0.892, which is significantly lower.

Let's visualize influence of *min_difference* attack parameter onto the attack result. The algorithm step has been increased for the effect to be more pronounced. In our case 99% of pixel weights lie in range [−1;1], so the difference by modulo is within [0;2], that's the range for the *min_difference* parameter. Fig. 5 has the results of targeted attack $0 \rightarrow 8$ displayed with algorithm parameters: $\alpha = 0.5$, *min_difference* = 1.2, *step* = 0.1, *max_steps* = 10. As it turns out, it has been enough to modify pixel brightness of only 4 source image points for the attack to succeed. SSIM value of 0.954 has been achieved on the 8th algorithm step. Experiments akin to the one presented above have been performed on a set of images and all with a success. However, it has been noted, that an algorithm has an interesting feature, where in some cases it leads the attack not directly to the target clas For example, for the targeted attack 6 into 1 during the first algorithm steps we have 6 misclassified as 2. In a case when such an intermediate class appears during the attack, perceptual image quality is degraded (for the example above we got SSIM = 0.635).

The problem's roots have been investigated by utilizing PCA (Principal Component Analysis). It has been noticed after having visualized scatted plot for points of classes 2, 6 and 1, that intercluster difference for digits 2 and 6 is a lot lower that the one for 6 and 1. And as is shown on fig. 6 (for α=0.5), the difference vector between the target and source classes passes through the field of twos.s, but through some intermediate instead.

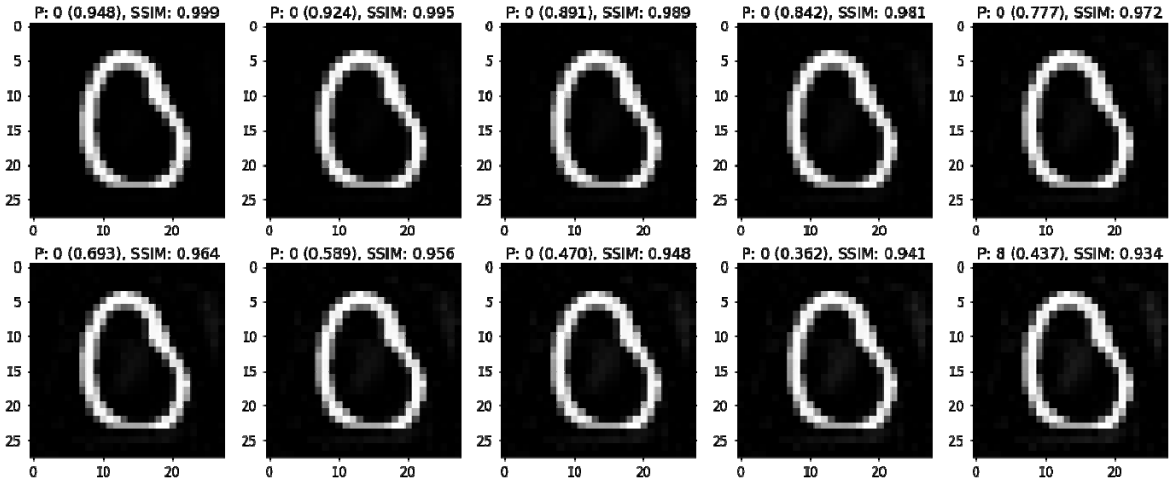

Figure 4 – $0 \rightarrow 8$ attack. Algorithm parameters: *min_difference* = 0.0, *step* = 0.01, max_*steps* = 10

DOI 10.15588/1607-3274-2019-2-12

While performing targeted attack such issues can be avoided by using another algorithm modification with parameter α=1. Thus, by applying such a modification on the one hand, for a targeted attack we will strive to maximize output of the target neuron when compared to others. On the other hand, we will minimize source neuron output in respect to other ones to accomplish non-targeted attack, which can be achieved with an additional algorithm modification with $\alpha = 0$.

Much better modified image quality can be obtained by the virtue of such algorithm modifications (for example, for the same $6 \rightarrow 1$ attack SSIM score has risen to 0.780). Fig. 6 has trajectories of the source image 6 while being a subject to modification by the original algorithm and both its variations (targeted and non-targeted).

5 RESULTS

Generalized attack analysis was performed next. By launching targeted attack for each pair of source and target classes, a success rate heatmap has been drawn (Fig. 7). Source classes are shown on the left, the target ones in the bottom.

Heatmap elements are SSIM values averaged across all the attacks for a given source, target pair. Mean quality over the whole test dataset is 0.76 – such images after attack will still be correctly classified by a human. Top score has been achieved for an attack of similar digits i.e.: $8 \rightarrow 9$, $9 \rightarrow 8$, $0 \rightarrow 8$, $3 \rightarrow 8$. The worst quality degradation was for attack $1 \rightarrow 0$. This can be explained by the fact that the vital region for zero is a black hole in the middle, which gets usually overlapped by a white bar of a one digit. Should be noted, that $0 \rightarrow 1$ attack requires much fewer image modifications then the one into opposite direction, which is proved by comparing SSIM metric value (higher by 0.21). As attacks have been built on real test set samples opposed to the generic digit silhouettes which got learnt by the neural network, the heatmap SSIM values lack symmetry. Training set quality loss heatmap has a similar feature in it.

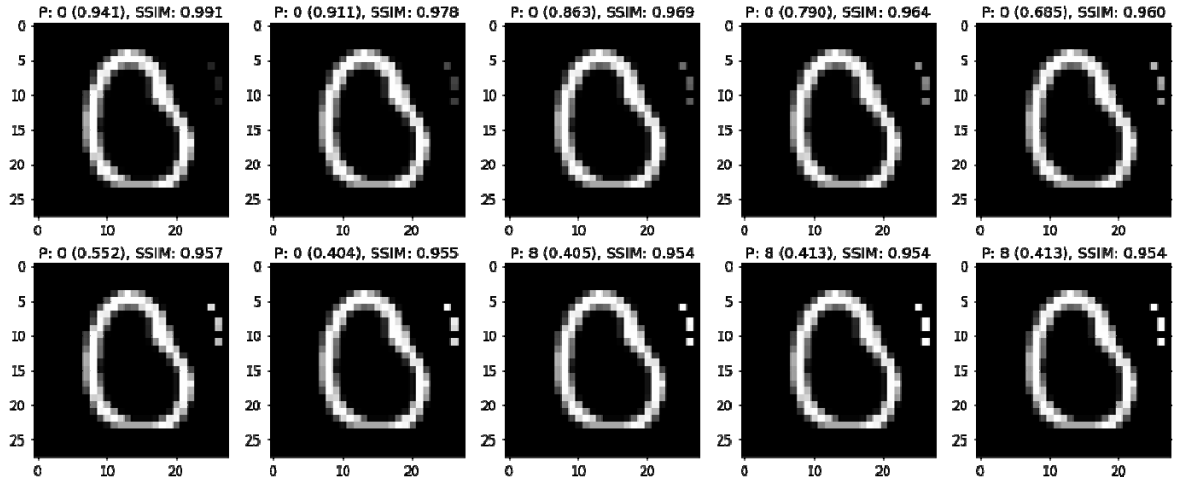

Figure 5 – $0 \rightarrow$ 8 attack**.** Algorithm parameters: *min_difference* = 1.2, *step* = 0.1, *max_steps* = 10

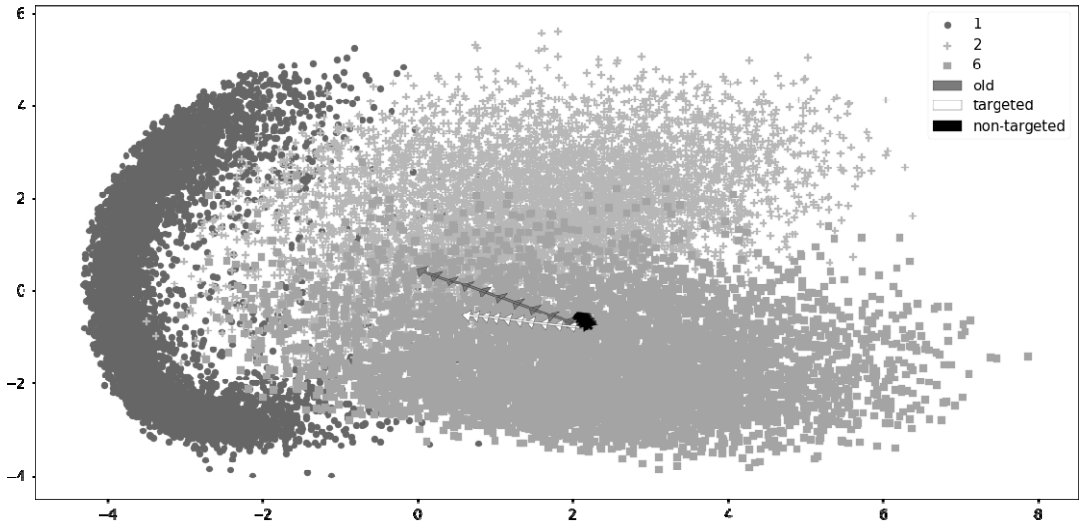

Figure 6 – $6 \rightarrow$ 1 attack trajectory as projected onto a surface using PCA: targeted attack with $\alpha = 0.5$ (the longest trajectory), targeted with α=1 (route of an average length), non-targeted with $\alpha = 0$ (the shortest path)

DOI 10.15588/1607-3274-2019-2-12

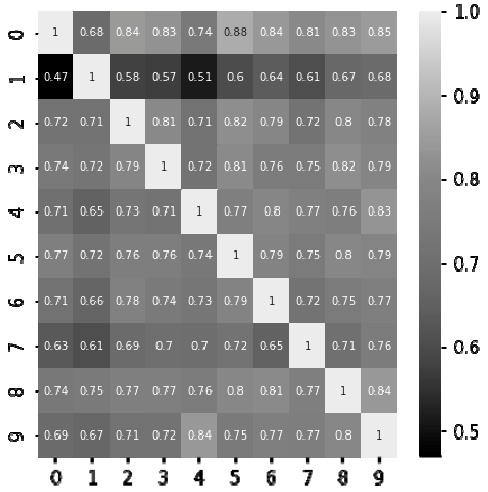

Figure 7 – SSIM metric values heatmap for each source, target attack pair

Lower average image quality loss can be attained by employing a stricter parameter selection algorithm. While fig. 7 has losses computed for a low empirically chosen *min_difference*=0.5 value, by selecting the best value from range [0.5, 1.2] an increase of SSIM to 0.87 score has been observed.

A higher average SSIM score 0.93 has been reached for the non-targeted attack case, which means that source and target images are nearly impossible to distinguish with a naked eye. As previously, digits 0, 8, were the best ones to attack, 1 has proved to be the most problematic (see the boxplot on Fig. 8). By analyzing non-targeted attack results, one can come up to a conclusion that if image quality loss deviation is high for different images of a certain class (i.e. some images are easy to attack, while others not), then algorithm is inefficient (as it can only change digits that look similar to several classes); if, conversely, the deviation is small, then the algorithm is efficient.

SSIM plots with respect to min_*difference* parameter value have allowed to make a conclusion about the fact that each class has a tendency of an image quality rise jointly with min_*difference* increase up to 0.9 point, such a trend is especially noticeable for the class of nines. Specifically, the human perceived image quality loss will be substantially lower in case of a strong change of several pixels, then when all image points are slightly modified. Taking this feature into account is the thing that makes our algorithm standout among all other known in literature methods.

Among the fast gradient methods, the most efficient algorithm is I-FGM with L_2 norm loss [7]. An attack for each source, target pair has been conducted by following the above described procedure for the case without *min_difference* selection. Algorithm has been successful on all test images but has achieved a lower SSIM score of 0.83.

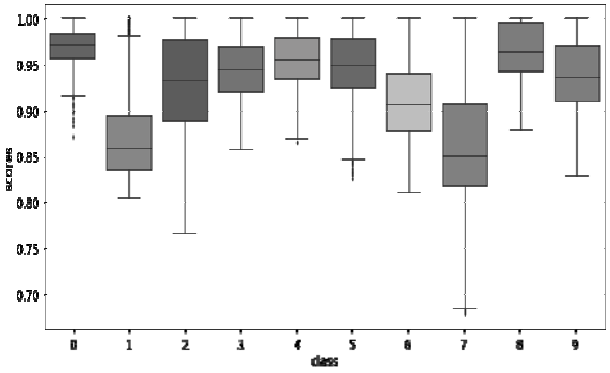

Figure 8 – Low non-targeted attack SSIM score deviation. Algorithm is good at attacking diverse images of a single class

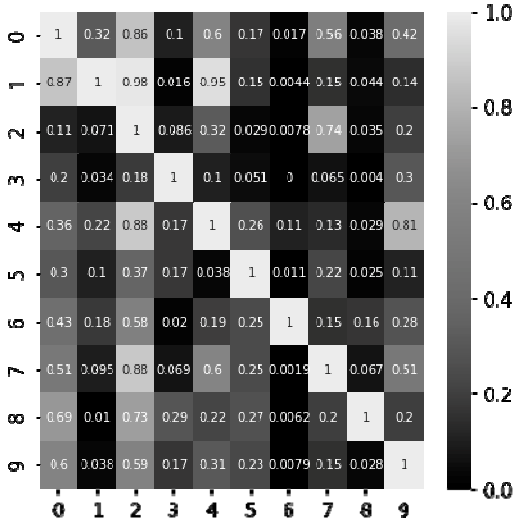

Figure 9 – Single-layer to 5 layer fully connected network transfer success probability heatmap

Lastly, the question of adversarial image transfer has been considered. This way, we want to perform the so-called black box attack, when we don't have any knowledge about network's weights or architecture, the only allowed operation is to query neural network prediction engine by submitting some images. The attack will be performed by using the above described logistic regression architecture, then an attempt to transfer each image to a 5-layered unknown neural network will be made.

For the results reproducibility neural network architecture is to follow, yet this knowledge has not been used in any way during the attack phase. The neural network has a 5-layer fully-connected architecture with layer sizes of 200, 100, 60, 30, i.e. 4 hidden, one output with 10 neurons and one input with 784. As a mean of regularizing the network Batch Normalization has been applied after the first layer, Dropout after the second one. ReLU has been used as an activation function for all layers but the last one, where we have switched to a Softmax function instead. After 100 epochs of training

DOI 10.15588/1607-3274-2019-2-12

using Adam optimization algorithm, training set accuracy has reached 98.65%, the test one 98.51%.

Figure 9 has a generalized heatmap representation of an attack transfer success probabilities. By the above-described procedure an average probability of 33% successfully transferred images has been achieved. Interesting to note, that in many cases images that were difficult to attack for the original network have seen a higher transfer rates than the ones needed only minor image changes. For instance, it has been possible to successfully transfer 87% of $1 \rightarrow 0$ attack images, which have been one of the most challenging ones, but only 14% $9 \rightarrow 7$ attack images.

Let's follow along the $4 \rightarrow 9$ attack procedure. Each step will have the predicted digit with its probability shown for the attacked single-layer classifier (SL) and 5-layer fully-connected network (FC5) (Fig. 10). It should be observed that after there were enough changes to cheat the original network, it has been necessary to make 4 more steps to deceive the 5-layer one. This means that while the two networks have a similar decision boundary yet each one has it biased with respect to another one.

Considering the above-described thoughts, a generalized targeted attack with neural network transfer has been conducted once again. This way, after having performed a successful attack on the source network, 5 more algorithm steps were made (where number 5 was chosen empirically), which makes it possible to transfer 91% of adversarial images with a minimal image quality loss.

6 DISCUSSION

Hence, another way of analyzing neural network safety has been presented (logistic regression in particular) against input data attacks. Simplified targeted and non-targeted logistic regression attack algorithms for handwritten digit classification problem on the MNIST dataset has been built. A visual "importance" interpretation of each image pixel for its classification as of an instance of a class has been given. An analysis has been performed that has permitted to define classes the most vulnerable to the attack as well as images for which class predicted by the neural network can be changed unnoticeably for a human being. The proposed algorithm gives a possibility of conducting a successful adversarial attack by modifying only several image pixels which minimizes image data loss.

The analysis of image quality loss performed based on Structural Image Similarity Index (SSIM) for targeted and non-targeted neural network attacks for the proposed algorithm shows that the developed algorithm provides a better image quality of the attack in comparison to other gradient methods.

Adversarial examples, built with the developed algorithm, have been successfully transferred to a different neural network with 5 layers of an unknown architecture. High change of adversarial image transfer to a network with a vastly different network architecture makes the algorithm applicable for attacking restricted-access systems.

CONCLUSIONS

A logistic regression adversarial attack algorithm for the MNIST dataset handwritten digit classification task has been proposed. Targeted and non-targeted neural network attacks can be performed by utilizing one of the two algorithm modifications. The relevance is explained by the fact of an increasing growth of neural network use in the field of public safety and of a critical need of exploring neural network attack methods and of their precursors.

The scientific novelty of obtained results is that for the first time adversarial attack algorithm has been built upon the attack scoping idea. The presented fast and efficient attack algorithm is able to attack both the whole image as well as separate image regions, which makes the attack algorithm more flexible. An image information loss can be minimized by modifying only a couple of pixels.

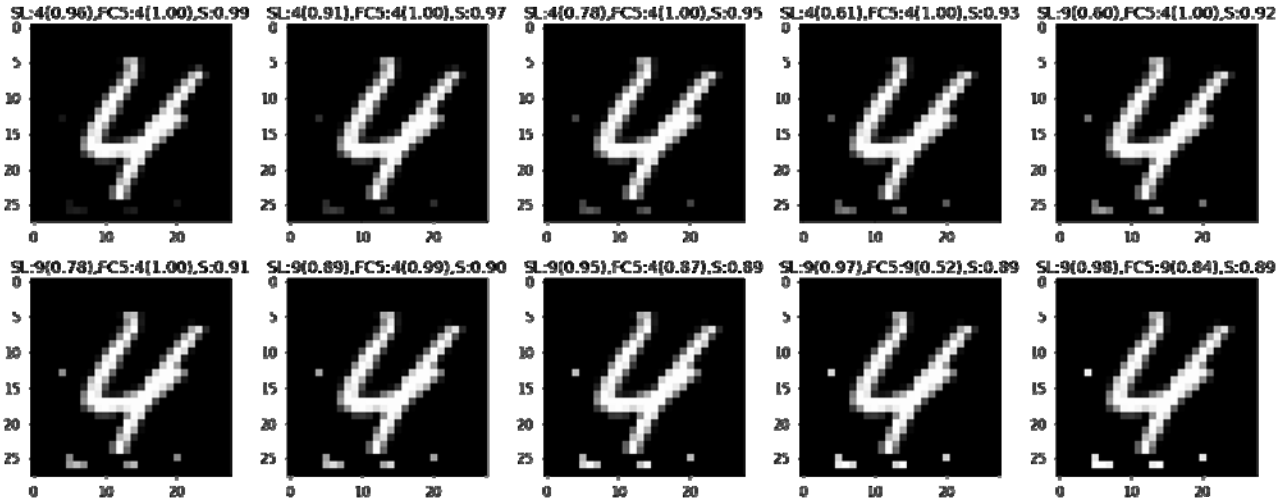

Figure 10 – $4 \rightarrow 9$ attack**.** Adversarial image transfer to a fully-connected 5-layer network. Algorithm parameters: min_*difference* = 0.45, *step* = 0.2, max_*steps* = 10

DOI 10.15588/1607-3274-2019-2-12

The practical significance of obtained results is that an early neural network vulnerability diagnostic can be performed by utilizing the proposed algorithms and image quality loss analysis system, which is a pivotal point towards a safer practical neural network use.

Prospects for further research are to study physical neural network adversarial attack transfer with the use of an ordinary pen, based on a pixel importance of the selected class.

ACKNOWLEDGMENTS

The authors would like to thank Doctor of Sciences in Physics and Mathematics, Associate Professor of the Department of Data Analysis and Artificial Intelligence of the National Research University "High School of Economics" Gromov Vasilii Aleksandrovich for support and a fruitful paper discussion.

REFERENCES

1. Krizhevsky A., Sutskever I., Hinton G. E. ImageNet classification with deep convolutional neural networks, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems : 3–6 December 2012: proceedings, Lake Tahoe.* Nevada, USA, NIPS, 2012, pp. 1106–1114. DOI:10.1145/3065386
2. Pang Wei Koh, Percy Liang Understanding Black-box Predictions via Influence Functions [Electronic resource]. Access mode: http://arXiv:1703.04730.
3. Szegedy C., Zaremba W., Sutskever I. et al. Intriguing properties of neural networks [Electronic resource]. Access mode: http://arXiv:1312.6199.
4. Zeiler M. D., Fergus Rob eds.: Fleet D., Pajdla T., Schiele B., Tuytelaars T. Computer Vision Visualizing and Understanding Convolutional Networks – ECCV 2014. Lecture Notes in Computer Science. Springer, Cham, 2014, Part 1, Vol 8689, pp. 818–833. DOI:10.1007/978-3-319-10590-1_53
5. LeCun Y. Corinna Cortes, Christopher J. C. Burges. The MNIST database of handwritten digits [Electronic resource]. Access mode: http://yann.lecun.com/exdb/mnist/.
6. Goodfellow Ian J., Shlens J., Szegedy C. Explaining and Harnessing Adversarial Examples [Electronic resource]. Access mode: http://arXiv:1412.6572.
7. Kurakin A., Goodfellow Ian J., Bengio S. Adversarial machine learning at scale [Electronic resource]. Access mode: http://arXiv preprint arXiv:1611.01236.
8. Dong Y., Liao F., Pang T. et al. Boosting adversarial attacks with momentum. [Electronic resource]. Access mode: http://arXiv:1710.06081.
9. Eykholt K., Evtimov I., Fernandes E. et al. Robust Physical-World Attacks on Deep Learning Models. [Electronic resource]. Access mode: http://arXiv preprint arXiv:1707.08945v5.
10. Kurakin A., Goodfellow I., Bengio S. Adversarial examples in the physical world [Electronic resource]. Access mode: http://arXiv preprint arXiv:1607.02533.
11. Sharif M., Bhagavatula S., Bauer L., Reiter M. K. Accessorize to a crime: Real and stealthy attacks on state-of the-art face recognition, *Computer and Communications Security: ACM SIGSAC Conference, Vienna, Austria,* 24–28 October 2016, proceedings. ACM, 2016, pp. 1528–1540. DOI: 10.1145/2976749.2978392
12. Chen Pin-Yu, Huan Zhang, Yash Sharma et al. ZOO: Zeroth Order Optimization based Black-box Attacks to Deep Neural Networks without Training Substitute Models, *Artificial Intelligence and Security: the 10th ACM Workshop AISec'17, Dallas, TX, USA, 30 October – 03 November 2017: proceedings. ACM New York, NY, USA, 2017*, pp. 15–26. DOI: 10.1145/3128572.3140448
13. Carlini N., Wagner D. Towards evaluating the robustness of neural networks [Electronic resource], Access mode: http://arXiv:1608.04644 [cs.CR].
14. Papernot N., Faghri F., Carlini N. et al. Technical Report on the CleverHans v2.1.0 Adversarial Examples Library [Electronic resource]. Access mode: http://arxiv.org/abs/1610.00768.
15. Xiaoyong Yuan, Pan He, Qile Zhu et al. Adversarial Examples: Attacks and Defenses for Deep Learning [Electronic resource]. Access mode: http://arXiv:1712.07107v2 [cs.LG].
16. Naveed Akhtar, Ajmal Mian. Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey [Electronic resource]. Access mode: http://arXiv:1801.00553 [cs.CV].
17. Papernot N., McDaniel P., and Goodfellow I. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples [Electronic resource]. Access mode: http://arXiv preprint arXiv:1605.07277.
18. Wang Z., Bovik A. C., Sheikh H. R., Simoncelli E. P. Image quality assessment: From error measurement to structural similarity, *IEEE Trans. Image Process,* 2004, Vol. 13, No. 4, pp. 600–612. DOI: 10.1109/tip.2003.819861

Received 02.11.2018.
Accepted 16.01.2019.

УДК 004.93

ПОКРАЩЕННЯ ЯКОСТІ ЗМАГАЛЬНОЇ АТАКИ ШЛЯХОМ УТОЧНЕННЯ ЇЇ ОБЛАСТІ

Хабарлак К. С. – магістр кафедри системного аналізу і управління Національного ТУ «Дніпровська політехніка», Україна.

Коряшкіна Л. С. – канд. фіз.-мат. наук, доцент кафедри системного аналізу і управління Національного ТУ «Дніпровська політехніка», Україна.

АНОТАЦІЯ

Актуальність. Предметом дослідження даної роботи є змагальні атаки, види, причини виникнення, а також алгоритми атак. Представлений швидкий спрощений і більш ефективний (порівняно з існуючими аналогами) алгоритм атаки на модель логістичної регресії. Актуальність роботи пояснюється малою дослідженістю критичної уразливості нейронних мереж – так званих змагальних прикладів, які дозволяють зламувати механізм передбачення і отримувати довільний результат, роблячи системи безпеки, засновані на нейронних мережах, малоефективними.

DOI 10.15588/1607-3274-2019-2-12

Мета. Розробка алгоритмів різних типів атаки на навчену одношарову нейронну мережу з урахуванням результатів попереднього аналізу параметрів самої мережі, а також оцінка втрат якості зображень, що були піддані модифікації, порівняння результатів проведення атак за допомогою розроблених алгоритмів і змагальних атак подібного роду.

Методи. На основі результатів аналізу матриць ваг навченої нейронної мережі сформульована ідея побудови алгоритмів атаки на нейронну мережу, виділяючи для атаки певні області на зображенні з урахуванням різниці вагових матриць цільового і вихідного класів. Представлений швидкий і досить ефективний алгоритм атаки, який здатний використовувати для атаки як все зображення повністю, так і окремі його регіони, що робить алгоритм більш гнучким. Використовуючи метрику структурної схожості зображень SSIM, проведений аналіз алгоритму і його модифікацій, а також порівняння його з попередніми методами, які використовують для атаки звичайний градієнт.

Результати. Побудовано спрощені алгоритми націленої і ненаціленої атак на одношарову нейронну мережу, яка застосовується для класифікації рукописних цифр з набору даних MNIST. Дана візуальна і змістовна інтерпретація налаштованих ваг мережі як «важливостей» точок зображення для розпізнавання його як представника того чи іншого класу. На основі порівняння структурної схожості зображень алгоритмом SSIM був проведений аналіз втрат якості зображень для задач націленої і ненаціленої атак наведеними спрощеними алгоритмами на всій тестової вибірці. Подібний аналіз дозволив визначити класи, що найбільш піддаються атакам, а також зображення, для яких клас, передбачений нейронною мережею, може бути змінений непомітно для людини.

Змагальні приклади, побудовані за допомогою розробленого в статті алгоритму, перенесені на мережу з 5-ю шарами невідомої архітектури. У ряді випадків зображення для класів, які було складно атакувати для вихідної мережі, вдалося перенести з більшим успіхом, ніж ті, для зміни класу яких було досить мінімальних змін.

Висновки. Побудовані на основі ідеї обмеження області атаки змагальні приклади, а також система (методика) аналізу вхідних даних легко узагальнюється і на інші задачі розпізнавання, що робить представлену методику придатною для аналізу ряду практичних задач. Отже, представлений ще один підхід до аналізу безпеки нейронних мереж (зокрема, логістичної регресії) проти атак на вхідні дані.

КЛЮЧОВІ СЛОВА: змагальні атаки, швидкий алгоритм змагальної атаки, логістична регресія, уразливість нейронних мереж.

УДК 004.93

УЛУЧШЕНИЕ КАЧЕСТВА СОСТЯЗАТЕЛЬНОЙ АТАКИ ПУТЕМ УТОЧНЕНИЕ ЕЕ ОБЛАСТИ

Хабарлак К. С. – магистр кафедры системного анализа и управления Национального ТУ «Днепровская политехника», Украина.

Коряшкина Л. С. – канд. физ.-мат. наук, доцент кафедры системного анализа и управления Национального ТУ «Днепровская политехника», Украина.

АННОТАЦИЯ

Актуальность. Предметом исследования данной работы являются состязательные атаки, виды, причины возникновения, а также алгоритмы атак. Представлен быстрый упрощенный и более эффективный (по сравнению с существующими аналогами) алгоритм атаки на модель логистической регрессии. Актуальность работы объясняется малой исследованностью критической уязвимости нейронных сетей – так называемых состязательных примеров, которые позволяют взламывать механизм предсказания и получать произвольный результат, делая системы безопасности, основанные на нейронных сетях, малоэффективными.

Цель. Цель данной работы – разработка алгоритмов разных типов атаки на обученную однослойную нейронную сеть с учетом результатов предварительного анализа параметров самой сети, а также оценка потерь качества изображений, подвергнутых модификации, сравнение результатов проведения атак с помощью разработанных алгоритмов и состязательных атак подобного рода.

Методы. На основе результатов анализа матриц весов обученной нейронной сети сформулирована идея построения алгоритмов атаки на нейронную сеть, выделяя для атаки определенные области на изображении с учетом разности матрицы весов целевого и исходного классов. Представлен быстрый и достаточно эффективный алгоритм атаки, который способен использовать для атаки как все изображение, так и отдельные его регионы, что делает алгоритм более гибким. Используя метрику структурной похожести изображений SSIM, был проведен анализ алгоритма и его модификаций, а также сравнение его с предыдущими методами, использующими для атаки обычный градиент.

Результаты. Построены упрощенные алгоритмы нацеленной и ненацеленной атак на однослойную нейронную сеть, которая применяется для задачи классификации рукописных цифр из набора данных MNIST. Дана визуальная и содержательная интерпретация настроенных весов сети как «важностей» точек изображения для распознавания его как представителя того или иного класса. На основе сравнения структурной похожести изображений алгоритмом SSIM был проведен анализ потерь качества изображений для задач нацеленной и ненацеленной атак приведенными упрощенными алгоритмами на нейронные сети на всей тестовой выборке. Подобный анализ позволил определить классы, наиболее подверженные атаке, а также изображения, для которых класс, предсказанный нейронной сетью, может быть изменен незаметно для человека.

Состязательные примеры, построенные с помощью разработанного в статье алгоритма, перенесены на сеть с 5-ю слоями неизвестной архитектуры. В ряде случаев изображения для классов, которые было сложно атаковать для исходной сети, удалось перенести с большим успехом, чем те, для изменения класса которых было достаточно минимальных изменений.

Выводы. Состязательные примеры, построенные на основе идеи ограничения области атаки, а также методику анализа входных данных легко обобщается и на другие задачи распознавания, что делает ее применимой для анализа ряда

DOI 10.15588/1607-3274-2019-2-12

практических задач. Таким образом, представлен еще один подход к анализу безопасности нейронных сетей (в частности, логистических регрессоров) против атак на входные данные.

КЛЮЧЕВЫЕ СЛОВА: состязательные атаки, быстрый алгоритм состязательной атаки, логистическая регрессия, уязвимость нейронных сетей.

ЛІТЕРАТУРА / ЛИТЕРАТУРА

1. Krizhevsky A. ImageNet classification with deep convolutional neural networks / A. Krizhevsky, I. Sutskever, G. E. Hinton // Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems: 3–6 December 2012: proceedings. – Lake Tahoe, Nevada, USA: NIPS, 2012. – P. 1106–1114. DOI: 10.1145/3065386
2. Pang Wei Koh. Understanding Black-box Predictions via Influence Functions [Electronic resource] / Pang Wei Koh, Percy Liang. – Access mode: http://arXiv:1703.04730.
3. Szegedy C. Intriguing properties of neural networks [Electronic resource] / C. Szegedy, W. Zaremba, I. Sutskever et al. – Access mode: http://arXiv:1312.6199.
4. Zeiler M. D. Visualizing and Understanding Convolutional Networks / Matthew D Zeiler, Rob Fergus eds.: Fleet D., Pajdla T., Schiele B., Tuytelaars T. // Computer Vision – ECCV 2014. Lecture Notes in Computer Science. – Springer, Cham, 2014. – Part 1. – Vol 8689. – P. 818–833. – DOI:10.1007/978-3-319-10590-1_53
5. LeCun Y. The MNIST database of handwritten digits [Electronic resource] / Y. LeCun, Corinna Cortes, Christopher J. C. Burges. – Access mode: http://yann.lecun.com/exdb/mnist/.
6. Goodfellow Ian J. Explaining and Harnessing Adversarial Examples [Electronic resource] / Ian J Goodfellow, J. Shlens, C. Szegedy. – Access mode: http://arXiv:1412.6572.
7. Kurakin A. Adversarial machine learning at scale [Electronic resource] / A. Kurakin, Ian J. Goodfellow, S. Bengio. – Access mode: http://arXiv preprint arXiv:1611.01236.
8. Dong Y. Boosting adversarial attacks with momentum. [Electronic resource] / Y. Dong, F. Liao, T. Pang et al. – Access mode: http://arXiv:1710.06081.
9. Eykholt K. Robust Physical-World Attacks on Deep Learning Models. [Electronic resource] / K. Eykholt, I. Evtimov, E. Fernandes et al. – Access mode: http://arXiv preprint arXiv:1707.08945v5.
10. Kurakin A. Adversarial examples in the physical world [Electronic resource] / A. Kurakin, I. Goodfellow, S. Bengio. – Access mode: http://arXiv preprint arXiv:1607.02533.
11. Sharif M. Accessorize to a crime: Real and stealthy attacks on state-of the-art face recognition / M. Sharif, S. Bhagavatula, L. Bauer, M. K. Reiter // Computer and Communications Security: ACM SIGSAC Conference, Vienna, Austria, 24 –28 October 2016: proceedings. – ACM, 2016. – P. 1528–1540. DOI: 10.1145/2976749.2978392
12. Chen Pin-Yu. ZOO: Zeroth Order Optimization based Black-box Attacks to Deep Neural Networks without Training Substitute Models / Pin-Yu Chen, Huan Zhang, Yash Sharma et al. // Artificial Intelligence and Security: the 10th ACM Workshop AISec'17, Dallas, TX, USA, 30 October – 03 November 2017: proceedings. – ACM New York, NY, USA, 2017. – P. 15–26. – DOI: 10.1145/3128572.3140448
13. Carlini N. Towards evaluating the robustness of neural networks [Electronic resource] / N. Carlini, D. Wagner. – Access mode: http://arXiv:1608.04644 [cs.CR].
14. Papernot N. Technical Report on the CleverHans v2.1.0 Adversarial Examples Library [Electronic resource] / N. Papernot, F. Faghri, N. Carlini et al. – Access mode: http://arxiv.org/abs/1610.00768.
15. Xiaoyong Yuan. Adversarial Examples: Attacks and Defenses for Deep Learning [Electronic resource] / Yuan Xiaoyong, Pan He, Qile Zhu et al. – Access mode: http://arXiv:1712.07107v2 [cs.LG].
16. Naveed Akhtar. Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey [Electronic resource] / Naveed Akhtar, Ajmal Mian. – Access mode: http://arXiv:1801.00553 [cs.CV].
17. Papernot N. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples [Electronic resource] / N. Papernot, P. McDaniel, and I. Goodfellow. – Access mode: http://arXiv preprint arXiv:1605.07277.
18. Image quality assessment: From error measurement to structural similarity / [Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli] // IEEE Trans. Image Process. – 2004. – Vol. 13, No. 4. – P. 600–612. – DOI: 10.1109/tip.2003.819861

DOI 10.15588/1607-3274-2019-2-12