Methods for Defending Deep Neural Networks Against Adversarial Attacks

Vitali Himbitski
Belarusian State University
Minsk, Belarus
gimbitskyvitaly@gmail.com

Angelina Varashevich
Belarusian State University
Minsk, Belarus
varvashevichangelina@gmail.com

Vasili Kovalev
Minsk, Belarus

***Abstract*—In this paper, Gradient-based adversarial attack methods on classification neural networks for image processing are discussed. And also the archeitecture of neural network for defense against adversarial attacks is proposed. For the proposed neural network architecture, the dependence of the preprocessing quality of the attacked images on the training time and the number of trained parameters has been studied.**

***Keywords*—Deep Neural Networks (DNN), Adversarial Attack, Autoencoder, Convolutional Neural Network**

I. Introduction

In modern society, deep neural networks (DNNs) are widely used in various fields such as voice assistants, smart homes, security systems, and medical equipment. They face significant challenges, such as performing financial transactions upon voice assistant requests, autonomous vehicles participating in traffic, and medical equipment detecting diseases based on analysis and images. However, what happens if these systems fail to perform their intended tasks correctly? This could occur due to insufficient system testing or intentional interference. For example, based on data analysis, it may be recommended for almost everyone to visit a doctor, even if they are healthy. This would lead to overload in medical facilities and a decrease in attention from healthcare professionals.

Since DNNs can be easily attacked, it is important to consider the possibilities of preventing attacks on neural networks, especially in situations where the consequences of distorted operations can be extremely serious, such as informing a patient that they are healthy when they are not. Analytical work has been conducted to create protective mechanisms based on the fundamental principles of attacks and defense. The article "Intriguing Properties of Neural Networks" [1] proposed a process for creating adversarial input data in the field of image processing, known as the L-BFGS method. It also hypothesized the instability of DNNs to adversarial images due to their nonlinearity.

Later, the article "Explaining and Harnessing Adversarial Examples" [2] was published, which extensively discusses the problem of adversarial images and their impact on DNNs and linear models. The research concluded that the instability of DNNs is specifically due to the linear nature of their components. The article introduced the FSGM attack, which is computationally lighter compared to the previously proposed L-BFGS method. The influence of this attack on various networks, including linear models trained on the MNIST dataset for classification, was described. An alternative approach to data augmentation and improving resilience to adversarial images was also proposed — training the network on generated adversarial images.

The book "Reliability of Neural Networks: Strengthening AI's Resistance to Deception" [3] provides information about various attacks on artificial intelligence, describes the conceptual foundations of generating adversarial input data, and explains the computational methods used to create them. The book also presents arguments about protecting DNNs from attacks.

The article "Adversarial Attacks on Reinforcement Learning" [4] presents different attacks on DNNs as well as various defense methods against such attacks.

The structure of this article is as follows. Section 2 describes the datasets used in this research. Then, in Section 3, the method for generating adversarial attacks is described. Section 4 briefly outlines the adversarial training method. Finally, Section 5 describes the defense methods against adversarial attacks proposed in this work and analyzes the results of protecting DNNs using these methods.

II. Training Data

In this paper, adversarial attacks were performed on classification neural networks for image processing. The following input datasets were used for training:

- CT slices of people with corresponding classes: shoulders, heart, liver. The size of the dataset is 1801 and the classes are balanced. The images are proposed in RGB (red, green, blue) format with size 512*512.
- Histologic breast cancer image set composed of images from 2 different hospitals in the Netherlands. The set was used in the CAMELYON-2016 competition, and includes the corresponding classes: normal

(NRM), tumor (TUM), epithelial tissue (EPI). The dataset size is 18000 and the classes are balanced. The images are proposed in RGB (red, green, blue) format with a size of 256*256.

Figures 1 and 2 are an example of the contents of CT slice set and histology set, respectively.

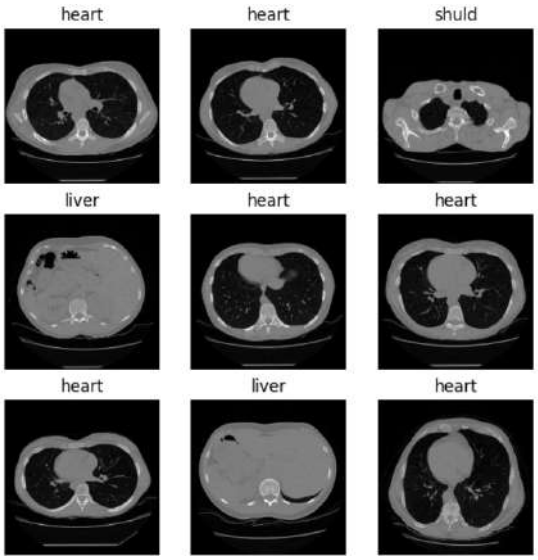

Figure 1. Example of images from the CT scan dataset.

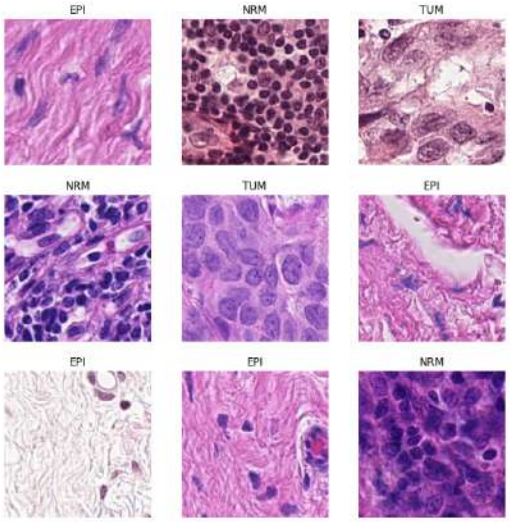

Figure 2. Example of images from a histology set.

III. Adversarial attaks

Let's consider an algorithm that classifies images. Let its processing be defined by the function:

$$f : X \to y \quad (1)$$

Here, XX is the set of images with a size equal to the input size of the classification algorithm, and yy is the set of numbers corresponding to class labels. Now, suppose there exists an attacking algorithm whose processing is defined by the function:

$$g : X \to X' \quad (2)$$

The sets of images XX and X'X' have different distributions, but the images in these sets have the same size.

If the attacking algorithm works effectively, then the following condition holds:

$$f(x) \neq f(g(x)), \quad x \in X \quad (3)$$

under the condition that:

$$\|x - g(x)\| < \epsilon \quad (4)$$

A. Gradient-based adversarial attacks

Gradient-based adversarial attacks are used to misclassify an image by a neural network. They can be directional and non-directional. Gradient-based adversarial attacks work as follows. The attacking algorithm generates noise from some distribution. The noise should have the same dimensions as the image itself. This noise is added to the image forming a new transformed image. The transformed image is fed to the input of the classification network. If the adversarial attack is successful, the neural network misclassifies the transformed image. Next, we will describe some of the most common methods of gradient attacks. They will also be used later in this paper.

B. Fast Gradient Sign Method (FGSM)

The Fast Gradient Sign Method (FGSM) **??** is an adversarial attack method that utilizes the gradients of a model to create adversarial examples. It was proposed in 2014 by researchers Ian Goodfellow, Jonathon Shlens, and Christian Szegedy.

The FGSM method defines the function (4) as:

$$g(x) = x + \epsilon \cdot \text{sign}(\nabla_x l(x, y_{\text{true}})) \quad (5)$$

Here, ϵ is the magnitude, and $l(x, y_{true})$ is the loss function of the classification neural network with respect to the true label.

The main idea of FGSM is to utilize the gradient information of the model's loss function with respect to the input data to create adversarial examples. Gradients indicate the direction in which the loss function will change the most when the input data is perturbed. The FGSM attacking algorithm utilizes this information to create adversarial examples that maximize the loss function, resulting in the misclassification of the image by the neural network.

The FGSM method, as presented in the implementation (5), can be used for untargeted attacks. Since this

implementation increases the loss function with respect to the true label, the classification neural network is more likely to predict the wrong class.

By using a different implementation of the attacking function in FGSM, a targeted attack can be achieved. To do this, it is sufficient to define the attacking function implementation as:

$$g(x) = x - \epsilon \cdot \text{sign}(\nabla_x l(x, y_{\text{target}})) \quad (6)$$

Here, ϵ is the magnitude, and $l(x, y_{target})$ is the loss function of the classification neural network with respect to the target label. This implementation minimizes the loss function of the classification neural network with respect to the target label, resulting in the neural network more likely to predict the target class for the given image.

C. Iterative Fast Gradient Sign Method (I-FGSM)

The Iterative Fast Gradient Sign Method (I-FGSM) is an adversarial attack method that is implemented by repeating the FGSM method a certain number of times. In I-FGSM, the attacking algorithm is implemented using the following functions:

$$g(x) = x_n \quad (7)$$

$$x_{t+1} = \text{clip}_\epsilon(x_t + \alpha \cdot \text{sign}(\nabla_x l(x_t, y_{\text{true}}))) \quad (8)$$

Here, $x_0 = x$, α is the step size for adapting x_t, $clip$ is a function that ensures $x_t \in (x-\epsilon, x+\epsilon)$, and $l(x_t, y_{true})$ is the loss function of the classification neural network with respect to the true label.

In the above implementation, the I-FGSM algorithm generates an untargeted adversarial attack. It increases the loss function of the classification neural network with respect to the true label, causing the neural network to more likely predict the wrong class for the given image.

This method can be modified to implement a targeted adversarial attack. For this, the attacking algorithm should be defined using the following function:

$$g(x) = x_n \quad (9)$$

$$x_{t+1} = \text{clip}_\epsilon(x_t - \alpha \cdot \text{sign}(\nabla_x l(x_t, y_{\text{target}}))) \quad (10)$$

Here, $l(x_t, y_{target})$ is the loss function of the classification neural network with respect to the target label. This implementation minimizes the loss function of the classification neural network with respect to the target label, resulting in the neural network more likely to predict the target class for the given image.

The I-FGSM attack method is stronger than FGSM. In experiments conducted in this study, it has been found that for the same magnitude of attacks, the I-FGSM method leads to a higher number of incorrect predictions by the considered neural networks compared to the FGSM method.

D. Generation of adversarial attacks

The experiment investigated the effect of FGSM on ResNet50, MobileNetV2 and DenseNet169 neural networks, which are widely used image classification models. Different attack scenarios were studied, including changing the input data, changing the noise magnitude ϵ, and applying defense techniques such as Adversarial Training and using an additional neural network for image preprocessing.

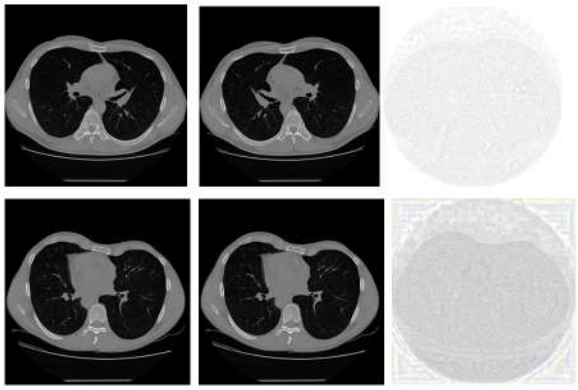

Figure 3. Example of adversarial attack when $\epsilon = 0.004$, (first and second rows, respectively). From left to right in the row — the original image, the malicious image, and the attack itself.

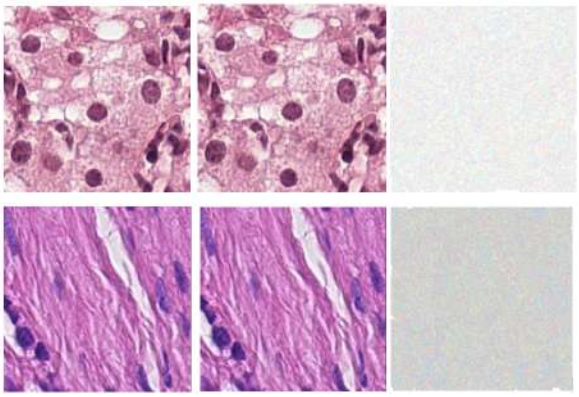

Figure 4. Example of adversarial attack when $\epsilon = 0.004$ (first and second rows, respectively). From left to right in the row — the original image, the malicious image, and the attack itself.

Figures 3, 4, 5 and 6 show examples of the original images, malicious images with added noise from the FGSM attack, and the attack itself. The values of each pixel have been magnified by a factor of 30 to visualize the differences between the images. In each figure, the first row shows the best classification quality result obtained at $\epsilon = 0.004$, while the second row shows the best result at $\epsilon = 0.01$. From left to right, each figure

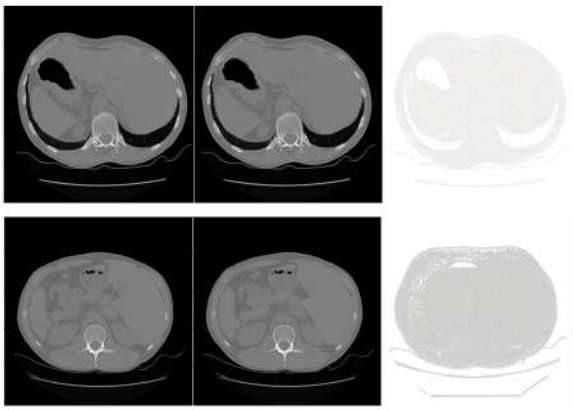

Figure 5. Example of adversarial attack when $\epsilon = 0.01$, (first and second rows, respectively). From left to right in the row — the original image, the malicious image, and the attack itself.

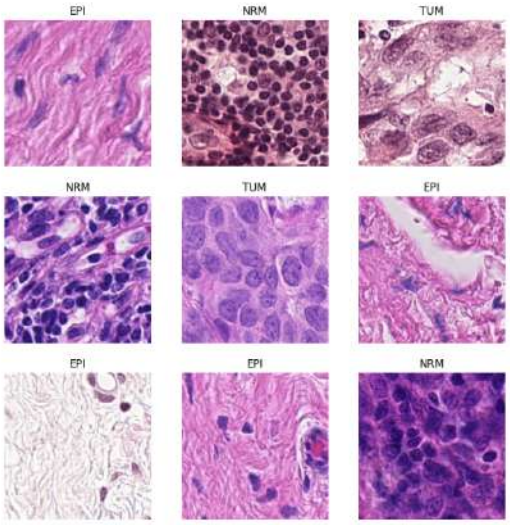

Figure 6. Example of adversarial attack when $\epsilon = 0.01$, (first and second rows, respectively). From left to right in the row — the original image, the malicious image, and the attack itself.

shows the original image, the malicious image, and the attack corresponding to that image.

E. Application of semantic technologies for image classification

Semantic approaches can also be used for the task of classifying attacked images. One such approach is to use semantic segmentation, which divides the image into semantic regions or objects. This allows the model to more accurately analyze the content of the image and classify it based on the different segments. Semantic segmentation can be achieved through a variety of techniques, including using convolutional neural networks with architectures specialized for segmentation, or combining deep learning with image processing techniques.

In addition, semantic techniques can be used to improve the interpretability of image classification. This is especially important in tasks where explanation and understanding of the decisions made are critical. Semantic technologies allow models to explain their classification decisions based on high-level concepts and semantic attributes. For example, one can use activation visualization techniques in neural networks to identify important image regions that influence classification, or to show relationships between classes and semantic concepts.

IV. Adversarial training

Adversarial training defense is as follows: the input dataset that learns to classify a DNN is augmented because the training dataset will now include adversarial images, a kind of augmentation. This form of data augmentation utilizes input data that is unlikely to be naturally acquired, but which reveals the shortcomings of the DNN. With the generated dataset, the model tries to minimize the loss function. In this way the DNN becomes less sensitive to small distortions. However, if malicious images produced by a different model but trained on a subset of the original input data, i. e., the model approximates the target DNN, are given as input, their behavior will be similar. That is, incorrect results may be obtained again, more or less different from the results obtained on the malicious images generated by the original network (from [2]).

In order to improve the robustness of the network using Adversarial training, the target DNN can be trained not only on malicious images generated by the original model, but also on other models. Since the model weights will be adjusted to the combination of malicious image groups, the DNN will become more robust to malicious images and it will be more difficult for an attacker to attack it.

V. Defense against adversarial attacks

In the Adversarial training procedure of the classification neural networks considered in this paper, we used the $SparseCategoricalCrossentropy$ loss function, $Adam$ optimizer with a training step of 10^{-3}. Since the classes in the input datasets are balanced, to assess the quality of the model we will use the accuracy quality metric, which characterizes the proportion of correct predictions relative to the number of all objects.

In this paper, a neural network with autoencoder architecture was used to implement the method of preprocessing attacked images. For the proposed neural network, the classification quality of the images preprocessed using it was measured.

The disadvantage of the autoencoder architecture is the significant reduction in the amount of information in the latent space relative to the amount of information of the image itself. To solve this problem, layers with information transfer from the previous state were added to the neural network architecture to protect against

adversarial attacks. The architecture of the proposed neural network is described in Table 3.

The neural network for image preprocessing was trained using a composite loss function consisting of image restoration loss function and classification loss function.

Also during the research of this image preprocessing method, the dependence of the classification quality of preprocessed images on the number of training epochs of neural networks for defense against adversarial attacks was measured.

The results of the conducted experiments are presented in Tables II and III.

From the results presented in the tables in Tables II and III, the following conclusions can be drawn:

1) The implementation of the attack depends on the choice of the dataset on which the target deep neural network (DNN) has been trained. When the same DNN is used to classify histology and computed tomography (CT) related datasets, the results after the FSGM attack do not match the accuracy of the original classification proportionally.
2) Attacks on different neural networks show different results. For example, the MobilenNetV2 neural network is found to be more resistant to attacks. This property is also manifested at different levels of attack magnitude. The results obtained depend primarily on the number of trained parameters in the neural network. The neural network with less number of trained parameters is less susceptible to attacks.
3) When the considered defense methods are applied on malicious images generated on the same networks, the classification accuracy on both datasets is recovered, and this recovery depends on the number of trained parameters. Although the MobileNetV2 neural network proved to be more resistant to attacks, its accuracy is recovered slower compared to the Resnet50 and Dencenet169 neural networks.
4) The amount of noise added to the original data to create malicious images affects the classification accuracy. The more noise added, the greater the deviation of the prediction vector from the original vector. Hence, the amount of noise required to create a malicious image depends on the particular network and the input dataset used.
5) The considered neural networks on a set of CT images showed high classification accuracy, indicating that the added noise is not enough to create malicious images.
6) When more noise is added to the original images at $\epsilon = 0.01$, the attack significantly reduces the accuracy of the classifier on the malicious image sets, except in the case of the CT image set for MobileNetV2, where the added noise at this value of ϵ was insufficient to turn the images into malicious images.

Thus, the classification accuracy of a DNN after applying FGSM to it is affected by the original input dataset, the architecture of the DNN, and the amount of noise superimposed on the images. The considered methods of protection against adversarial attacks showed high quality on all pairs of network-dataset type. This indicates the effectiveness of this type of protection to ensure the network resistance to FSGM attack.

VI. Conclusion

In this paper, a comprehensive analysis of gradient-based attacks is performed, taking into account the influence of the Deep Neural Network (DNN) architecture, the input dataset used, and the amount of added noise ϵ to form malicious images. Adversarial training and image preprocessing using neural networks were applied as defense techniques against the FGSM attack. These approaches achieved high classification accuracy on the original FGSMs, but did not fully recover the original accuracy level.

The results emphasize the importance of security techniques in deep neural networks. The study of attacks on such networks is of great importance for the development of effective defense mechanisms. The considered problematic requires further in-depth research, especially in light of the wide application of deep neural networks in various fields, including medicine, where classification accuracy is of particular importance and the potential consequences of errors are unacceptably high.

References

[1] Ch. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, "Intriguing properties of neural networks," arXiv preprint arXiv:1312.6199, 2013.

[2] I. Goodfellow, J. Shlens, Ch. Szegedy, Explaining and Harnessing Adversarial Examples, InternationalConference on Learning Representations, ICLR 2015, San Diego, CA, USA, 2015.

[3] Warr K, "Reliability of neural networks: strengthening AI robustness to deception," SPb: Peter, 2021, pp. 272.

[4] Weilin Xu, David Evans, Yanjun Qi, "Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks," arXiv preprint arXiv:1704.01155v2, 2017.

[5] Seungju Cho, Tae Joon Jun, Byungsoo Oh, Daeyoung Kim, "DAPAS : Denoising Autoencoder to Prevent Adversarial attack in Semantic Segmentation," arXiv preprint arXiv:1908.05195v4, 2020.

[6] Baoyuan Wu, Shaokui Wei, Mingli Zhu, Meixi Zheng, Zihao Zhu, Mingda Zhang, Hongrui Chen, Danni Yuan, Li Liu, Qingshan Liu, "Defenses in Adversarial Machine Learning: A Survey," arXiv preprint arXiv:2312.08890v1, 2023.

[7] Boxi Wu, Heng Pan, Li Shen, Jindong Gu, Shuai Zhao, Zhifeng Li, Deng Cai, Xiaofei He, Wei Liu, "Attacking Adversarial Attacks as A Defense," arXiv preprint arXiv:2106.04938v1, 2021.

[8] Ergezer M., Duong Ph., Green Ch., Nguyen T., Zeybey A., "A One Noise to Rule Them All: Multi-View Adversarial Attacks with Universal Perturbation," arXiv preprint arXiv:2404.02287, 2024.

Table I
Results of the experiments with $\epsilon = 0.004$

Data $\varepsilon = 0.004$	No attack	FSGM	Adversarial training	Preprocessing data	Pretrained
Histology	90.11	71.78	77.6	83.5	no
CT	100	100	100	100	no
Histology	94.65	60.73	91	93.6	yes
CT	100	86.12	100	100	yes

Table II
Results of the experiments with $\epsilon = 0.01$

Data $\varepsilon = 0.004$	No attack	FSGM	Adversarial training	Preprocessing data	Pretrained
Histology	90.11	36	83	85.5	no
CT	100	90	99.2	100	no
Histology	94.65	9.98	94.1	94.5	yes
CT	100	5.59	99.6	100	yes

Table III
Neural network architecture for image preprocessing

Layers	Input Shape $\rightarrow$ Output Shape	Layers Information
Input Layer	(1, 224, 224) $\rightarrow$ (32, 112, 112)	Conv2d
Encoder	(32, 112, 112) $\rightarrow$ (64, 56, 56)	Conv2d, LeakyReLU
	(64, 56, 56) $\rightarrow$ (128, 28, 28)	Conv2d, LeakyReLU
	(128, 28, 28) $\rightarrow$ (256, 14, 14)	Conv2d, LeakyReLU
Hidden Layer	(256, 14, 14) $\rightarrow$ (256, 14, 14)	Reshape, Linear, Reshape, Skip Connection
Decoder	(256, 14, 14) $\rightarrow$ (128, 28, 28)	Conv2dTransform, LeakyReLU, Skip Connection
	(128, 28, 28) $\rightarrow$ (64, 56, 56)	Conv2dTransform, LeakyReLU, Skip Connection
	(64, 56, 56) $\rightarrow$ (32, 112, 112)	Conv2dTransform, LeakyReLU, Skip Connection
Output Layer	(32, 112, 112) $\rightarrow$ (1, 224, 224)	Conv2dTransform, ReLU

[9] Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, Xueqi Cheng, "Multi-granular Adversarial Attacks against Black-box Neural Ranking Models," arXiv preprint arXiv:2404.01574, 2024.

[10] Amirian M., "Deep Learning for Robust and Explainable Models in Computer Vision," arXiv preprint arXiv:2403.18674v1, 2024.

МЕТОДЫ ЗАЩИТЫ ГЛУБОКИХ НЕЙРОННЫХ СЕТЕЙ ОТ ВРАЖДЕБНЫХ АТАК

Гимбицкий В. В., Варашевич А. Г., Ковалёв В. А.

В данной статье рассматриваются методы градиентной атаки на классифицирующие нейронные сети для обработки изображений. А также предложена архитектура нейронной сети для защиты от злоумышленных атак. Для предложенной архитектуры нейронной сети исследована зависимость качества предварительной обработки атакуемых изображений от времени обучения и количества обучаемых параметров.

Received 14.03.2024