ASSESSING THE VULNERABILITY OF AI-BASED SOLUTIONS IN HISTOPATHOLOGY OF CANCER

I. A. Filipovich, V. A. Kovalev

Belarusian State University, Minsk, Belarus
United Institute of Informatics Problems Belarus
National Academy of Sciences, Minsk, Belarus
E-mail: FilipovichIgor@yandex.by, vassili.kovalev@gmail.com

In this paper, we experimentally study the robustness of the Convolutional Neural Networks (CNNs) to adversarial attacks in different scenarios of computerized disease diagnosis. In order to disclose practically-relevant solutions, we attempt to compare the final CNN vulnerability scores under the condition of the use of different kinds of adversarial attacks as well as defense methods. On all occasions, we attempt to compare the basic and the most advanced solutions being available in every direction of the inquiry. In order to achieve this, we investigate EfficientNet CNN as one of the most popular convolutional networks. Also, we study the following three types of adversarial attacks: the FGSM Attacks, the Carlini-Wagner attacks as well as the AutoAttacks. After that, we examined three types of adversarial defenses including Adversarial Training, High-Level Representation Guided Denoiser, and the MagNet. The experiments have been performed on medical images typically used for computerized disease diagnosis in oncology (the whole-slide histology)

Keywords*: convolutional neural networks; adversarial attacks; biomedical images.*

Introduction. Deep neural networks are becoming more and more powerful machine learning tools. DNN can be applied to various areas of life: computer vision, sound and video processing, natural language processing. However, despite the ability of neural networks to show incredible results, they are not a universal solution. In addition, due to the strong dependence of neural networks on the quality and volume of the training sample, such models are unstable to disturbances in the input data. Moreover, when the task has a higher degree of responsibility, such as medical problems, the importance of robustness of the model to adversarial examples cannot be overestimated.

That's why we need to investigate the influence of different adversarial attacks on various medical images and what is more important we try to find the way to protect classification model from that attacks. As a baseline solution for model defense we study Adversarial Training and compare it to more complex defenses based on neural network autoencoders.

Materials. In this paper, we consider the dataset of the whole-slide histology. Dataset consists of four classes: ovary norm, ovary tumor, thyroid norm, thyroid tumor.

Methods. In our experiments, we performed three types of adversarial attacks:

1. FGSM attack [3] – the fast gradient sign method, where the perturba

tion noise is denoted by the following equation: $\varepsilon sign\nabla_x J(\theta, x, y)$.

2. AutoAttack [2] – a parameter-free, computationally affordable and user-independent ensemble of complementary attacks to estimate adversarial robustness.
3. Carlini-Wagner attack [1] – attack algorithms that was developed to show the weaknesses of defense methods and which are successful on both distilated and undistilated neural networks with 100% probability.

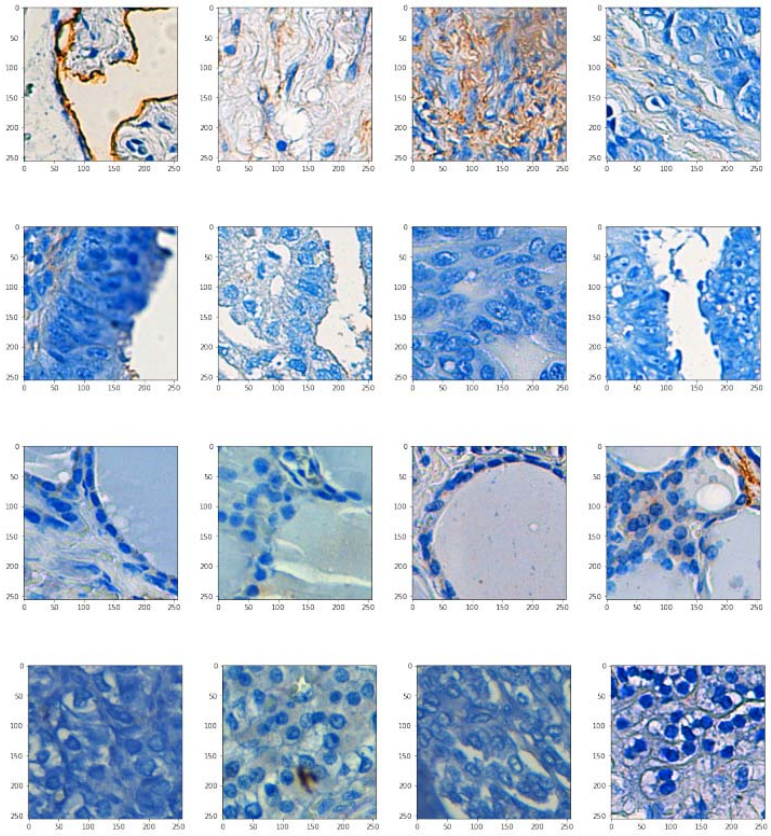

Fig 1. Histo dataset images samples. Each line corresponds to one class.

To prevent the influence of that attacks we tried three types of defenses:

Adversarial Training – means adding adversarial examples into training dataset and fine-tuning model on both attacked and clear images.

High-Level Representation Guided Denoiser [4] (Class Label Guided version) – denoising UNET, that is trained on classification problem loss.

MagNet [5] – one or more separate detector networks and a reformer network.

The whole pipeline consists of 5 stages:

1. We train classifier based on pretrained EfficientNet B3 on clean images from the described dataset and check its accuracy on specially prepared test set.

We check classifier accuracy against each specified above attack.

For each attack (FGSM, AutoAttack CW attack) independently we perform adversarial training. Then, for each classifier we examine accuracy against corresponding attacks and compare the effectiveness of adversarial training in relation to attacks.

Again, for each attack independently we train Class Label Guided Denoiser and check the robustness of ensemble of CGD and trained classifier from the first step.

We train MagNet autoencoders on clean dataset and perform attacks on the ensemble of MagNet and classifier from the first step.

Results. Experiments described above were carried for histologies dataset described in materials section. Results of these experiments are presented in Table 1.

Table 1

Histologies dataset accuracy table

	No defense	**Adversarial Training**	**CGD**	**MagNet**
No attack	95.8%	-	-	-
FGSM	17.8%	89.4%	91.4%	71.3%
Auto Attack	0%	93.1%	94.2%	72.7%
CW attack	10.2%	82.7%	92.9%	73.0%

Conclusions. In this study, we considered the vulnerability of EfficientNet B3 to adversarial attacks in an application to two different medical problems. Was discovered that model becomes powerless against adversarial examples, while the influence of adversarial noise cannot be seen by a human eye, what makes these attacks a real threat for deep neural networks, especially in a medical domain.

We should say that all considered defense methods protect the model to some extent, however all of them are not perfect and have their own advantages and disadvantages.

REFERENCES

1. Carlini, N., & Wagner, D. Towards evaluating the robustness of neural networks. IEEE Symposium on Security and Privacy (SP). pp. 39-57. 2017. DOI: 10.1109/SP.2017.49.
2. Croce, F., & Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. ICML. pp. 2206-2216. 2020.
3. Goodfellow, I., Shlens, J., & Szegedy, C. Explaining and harnessing adversarial examples. ICML. pp. 1-10. 2015.
4. Liao, F., Liang, M., Dong, Y., Pang, T., Hu, X., & Zhu, J. Defense against adversarial attacks using high-level representation guided denoiser. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1778-1787. 2018. DOI: 10.1109/CVPR.2018.00191
5. Meng, D., & Chen, H. Magnet: A two-pronged defense against adversarial examples. ACM SIGSAC Conference. pp. 135-147. 2017. DOI: 10.1145/3133956.3134057