5. Khosravan N., Bagci U. S4ND: Single-Shot Single-Scale Lung Nodule Detection // Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. Springer, Cham. 2018. Vol. 2. P. 794–802. DOI: https://doi.org/10.1007/978-3-030-00934-2_88
6. Benjdira B., Khursheed T., Koubaa A. et al. Car Detection using Unmanned Aerial Vehicles: Comparison between Faster R-CNN and YOLOv3 // Proceedings of the 1st International Conference on Unmanned Vehicle Systems (UVS). 2019. DOI: 10.1109/UVS.2019.8658300
7. Islam M. R., Shahid N., Karim D. T. et al. An efficient algorithm for detecting traffic congestion and a framework for smart traffic control system // Proceedings of the 18th International Conference on Advanced Communication Technology (ICACT). 2016. P. 802–807. DOI: 10.1109/ICACT.2016.7423566
8. Vasilev I., Slater D., Spacagn G. et al. Python Deep Learning: Exploring deep learning techniques and neural network architectures with PyTorch, Keras, and TensorFlow, 2nd Edition. Birmingham, UK: Packt Publishing Ltd. 2019.
9. Redmon J., Farhadi A. YOLOv3: An Incremental Improvement [Electronic resource]. – Mode of access: https://arxiv.org/abs/1804.02767 – Date of access: 17.03.2020.
10. Blue S. T., Brindha M. Edge detection based boundary box construction algorithm for improving the precision of object detection in YOLOv3 // 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT). 2019. DOI: 10.1109/ICCCNT45670.2019.8944852
11. Kim K., Kim P., Chung Y., Choi D. Performance Enhancement of YOLOv3 by Adding Prediction Layers with Spatial Pyramid Pooling for Vehicle Detection // 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). 2018. DOI: 10.1109/AVSS.2018.8639438
12. Choi J., Chun D., Kim H. et al. Gaussian YOLOv3: An Accurate and Fast Object Detector Using Localization Uncertainty for Autonomous Driving [Electronic resource]. – Mode of access: https://arxiv.org/abs/1904.04620. – Date of access:17.03.2020.

A COMPARATIVE STUDY OF WHITE-BOX AND BLACK-BOX ADVERSARIAL ATTACKS TO THE DEEP NEURAL NETWORKS WITH DIFFERENT ARCHITECTURES

D. M. Voynov, V. A. Kovalev

United Institute of Informatics Problems
Belarus National Academy of Sciences, Minsk, Belarus

E-mail: {voynovdd, vassili.kovalev}@gmail.com

A few years ago, it was discovered that the Deep Convolutional Neural Networks (CNN) are vulnerable to so-called adversarial attacks. An adversarial attack supposes a subtle modification of an original image in such a way that the changes are almost invisible to the human eye. In this work, we are concentrating on biomedical images, which are playing the key role in the disease diagnosis and monitoring of various treatment processes. We present detailed results on the success rate for both white-box and black-box untargeted attacks to five types of popular deep CNN architectures including InceptionV3, Xception, ResNet50, DenseNet121, and Mobilenet.

Key words: convolutional neural networks; adversarial attacks; biomedical images.

Introduction. Recently, the deep convolutional neural networks (deep CNNs) are widely used in different application domains. They have demonstrated high-quality results in a number of various image processing, image analysis, and image recognition tasks. One of the vital fields of application of related deep learning technologies is the medical diagnosis [1, 2] where corresponding neural network solutions are typically aimed at the biomedical image segmentation and classification. Once obtained, these classification results can be used for the generation of so-called "second opinion" to be considered by medical doctors who are making the final diagnostic decisions.

Unfortunately, a few years ago it was discovered that the deep CNNs are vulnerable to so-called adversarial attacks [3]. An adversarial attack supposes a subtle modification of an original image in such a way that the changes are almost invisible to the human eye. The modified image is called an adversarial (attacking) image, and when submitted to a classifier it is misclassified. Thus, the goal of modification of an input image is to fool the CNN and force it to make a wrong classification decision. The adversarial attacks are called the white-box attacks when the information about the architecture and all the weights of the target CNN is known to the attacker. Alternatively, in case no information about the target CNN is available the attack is termed as a black-box attack. Also, when the attacker wants to force CNN to categorize an image to a specific wrong class, such an attack is referred to as a targeted attack. Otherwise, when the goal is to force CNN to a wrong decision, no matter to which wrong class the adversarial image will be classified, such an attack called the untargeted one.

Currently, there are a number of papers dedicated to the problem of adversarial attacks (see, for example, [4-8]). However, the majority of these works studying various aspects of the problem of adversarial attacks to images belonging to the computer vision domain and even to such benchmarking image datasets as popular MNIST images of 10 digits.

In this work, we are concentrating on biomedical images, which are playing the key role in the disease diagnosis and monitoring of various treatment processes. It is clear that the security of computer-assisted diagnosis processes is of paramount importance. We present detailed results on the success rate for both white-box and black-box untargeted attacks to five types of popular deep CNN architectures including InceptionV3, Xception, ResNet50, DenseNet121, and Mobilenet. The image data being used are digital chest X-Ray images, 2D slices of Computed Tomography (CT) images as well as color histology images.

Materials. The image data used in this study are digital chest X-Ray images of norm of different age groups, 2D slices of Computed Tomography (CT)

images representing the norm and lung tuberculosis as well as color histology images sampled from normal and cancerous tissue of thyroid glands and the ovary. An additional benchmarking image dataset consisted of 6 classes of histological images stained with different histochemical markers. A detailed description of image data and the number of images in each class is given in table 1.

Table 1.

Description of datasets and classification task configurations

Image type	Classification task	Number of images, total	Number of images by classes
Chest X-ray of Norm	2 age groups: G1: 20-35 years G2: 50-70 years	200,000	G1: 100,000 G2: 100,000 G3: 183,360
Histology, Ovary cancer and cancer of Thyroid gland	4 classes: C1: Ovary norm C2: Ovary tumor C3: Thyroid norm C4: Thyroid tumor	192,000	C1: 48,000 C2: 48,000 C3: 48,000 C4: 48,000
Histology images stained with conventional H&E method and specific targeted markers including CD31, CD105, D240, FRES, and Ki67	6 classes: C1: CD31 C2: CD105 C3: D240 C4: FRES C5: H&E C6: Ki67	267,984	C1: 59,568 C2: 37,488 C3: 55,296 C4: 35,280 C5: 24,192 C6: 56,160
Lungs CT, 2D axial slices, layers	2 classes: C1: Norm C2: Tuberculosis	149,248	C1: 111,990 C2: 37,258

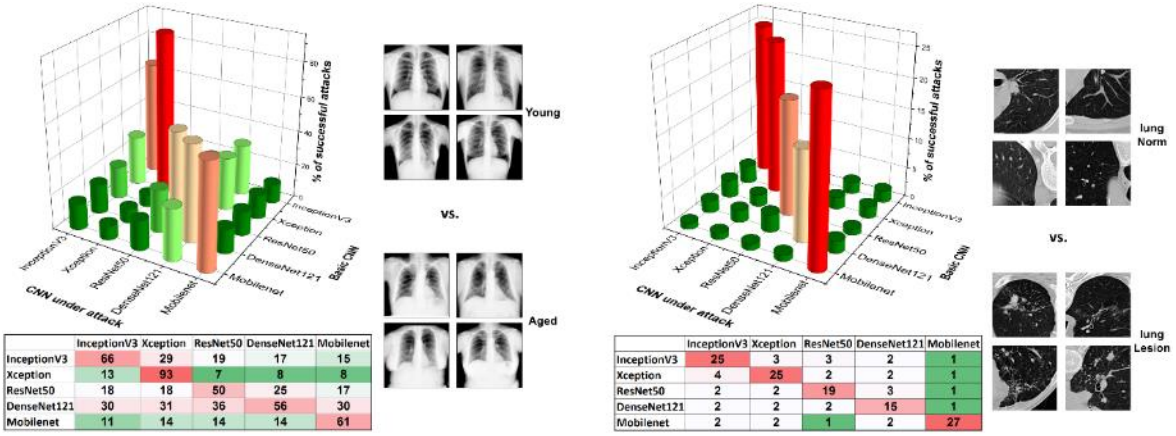
Methods. The black-box settings followed in this study require complete knowledge of the training image dataset of the network to be attacked (target CNN) while its architecture and trained parameters remain unknown. The whole attacking pipeline is based on the white-box Projected Gradient Descent (PGD) algorithm which was applied in the following way:

1. Train the “basic” CNN on which the adversarial (i.e., attacking) images will be generated using the training dataset of the target network being under attack.
2. Perform PGD attack on the trained network for each image in particular testing dataset to obtain a set of adversarial examples.
3. Pass both the testing dataset and its adversarial examples to the target network to assess the rate of successful attacks.

We used the rate of successful attacks as an estimation for impact of suggested algorithm. In order to make the results clear and consistent we define a single testing dataset for each training dataset. We have performed a large number of experiments based on each image dataset for the following five network architectures: InceptionV3, ResNet50, DenseNet121, Mobilenet, Xception. To imitate the black-box constraints the following testing pipeline was implemented:

1. Select a single network as a target one.
2. Perform the defined black-box algorithm with each network left as an attacking ones separately.
3. Carry out steps 1-2 with subsequent selection of every network as a target one.

Results. Experiments described above were carried out for every dataset described in the section of materials. Results of these experiments are presented in Fig. 1 and 2.

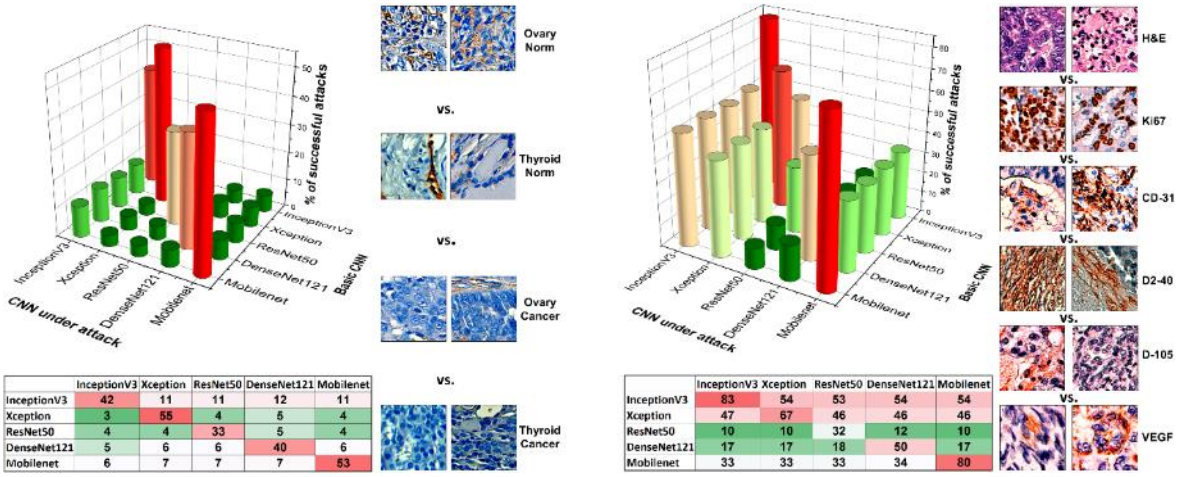


	InceptionV3	Xception	ResNet50	DenseNet121	Mobilenet
InceptionV3	66	29	19	17	15
Xception	13	93	7	8	8
ResNet50	18	18	50	25	17
DenseNet121	30	31	36	56	30
Mobilenet	11	14	14	14	61

	InceptionV3	Xception	ResNet50	DenseNet121	Mobilenet
InceptionV3	25	3	3	2	1
Xception	4	25	2	2	1
ResNet50	2	2	19	3	1
DenseNet121	2	2	2	15	1
Mobilenet	2	2	1	2	27

Fig.1. Results of adversarial attacks to X-Ray (left) and CT (right) images

The percentage of successful attacks is depicted both as the plots and corresponding data tables underneath. In order to ease interpretation of the results, the examples of original images of each class are provided on the right in each occasion. As it can be seen from the figures, every network is unstable under white-box attacking setup (see the bars on the leading diagonal). However, networks trained on CT and histology images of ovary and thyroid gland appear to be almost invulnerable to black-box attacks.



	InceptionV3	Xception	ResNet50	DenseNet121	Mobilenet
InceptionV3	42	11	11	12	11
Xception	3	55	4	5	4
ResNet50	4	4	33	5	4
DenseNet121	5	6	6	40	6
Mobilenet	6	7	7	7	53

	InceptionV3	Xception	ResNet50	DenseNet121	Mobilenet
InceptionV3	83	54	53	54	54
Xception	47	67	46	46	46
ResNet50	10	10	32	12	10
DenseNet121	17	17	18	50	17
Mobilenet	33	33	33	34	80

Fig.2. Results of adversarial attacks to histology images representing the Norm and Tumor classes of the ovary and thyroid gland (left) and 6 classes representing different histochemical markers (right)

Conclusions. Results obtained with this study allow drawing the following conclusions:

1. All tested networks are vulnerable to white-box adversarial attacks.
2. Network's vulnerability to black-box attacks strongly depends on the training dataset.

REFERENCES

1. Litjens G., Kooi T., Bejnordi B. et al. A survey on deep learning in medical image analysis // Med. Image Anal. 2017. Vol. 42. P. 60–88.
2. Ker J., Wang L., Rao J., Lim T. Deep Learning Applications in Medical Image Analysis // IEEE Access. 2018. Vol. 6. P. 9375–9389.
3. Szegedy C., Wojciech Z., Sutskever I. et al. Intriguing properties of neural networks // International Conference on Learning Representations (ICLR). 2014. P. 1–10.
4. Madry A., Makelov A., Schmidt L. et al. Towards Deep Learning Models Resistant to Adversarial Attacks [Electronic resource]. – Mode of access: https://arxiv.org/abs/1706.06083. – Date of access: 15.03.2020.
5. Xu W., Evans D., Qi Y. Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks [Electronic resource]. – Mode of access: https://arxiv.org/abs/1704.01155. – Date of access: 15.03.2020.
6. Wang H., Yu C. A Direct Approach to Robust Deep Learning Using Adversarial Networks [Electronic resource]. – Mode of access: https://arxiv.org/pdf/1905.09591. – Date of access: 13.03.2020.
7. Papernot N., McDaniel P., Jha S. et al. The Limitations of Deep Learning in Adversarial Settings [Electronic resource]. – Mode of access: https://arxiv.org/abs/1511.07528v1. – Date of access: 13.03.2020.
8. Sun K., Zhu Z., Lin Z. Towards Understanding Adversarial Examples Systematically: Exploring Data Size, Task and Model Factors [Electronic resource]. – Mode of access: https://arxiv.org/abs/1902.11019. – Date of access: 13.03.2020.