How Adversarial attacks affect Deep Neural Networks Detecting COVID-19?

Reza Amini Gougeh (✉ Amini.Gougeh.Reza@inrs.ca)

Institut National de La Recherche Scientifique (INRS-EMT) <https://orcid.org/0000-0002-1118-2941>

Short Report

Keywords: Adversarial Attacks, COVID-19, Medical Image Classification, Machine Learning Vulnerability

Posted Date: July 30th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-763355/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

How Adversarial attacks affect Deep Neural Networks Detecting COVID-19?

Reza Amini Gougeh¹

Abstract—Considering the global crisis of Coronavirus infection (COVID-19), the essence of utilizing novel approaches to achieve quick and accurate diagnosing methods is required. Deep Neural Networks (DNN) showed outstanding capabilities in classifying various data types, including medical images, in order to build a practical automatic diagnosing system. Therefore, DNNs can help the healthcare system to reduce patients waiting time. However, despite acceptable accuracy and low false-negative rate of DNNs in medical image classification, they have shown vulnerabilities in terms of adversarial attacks. Such input can lead the model to misclassification. This paper investigated the effect of these attacks on five commonly used neural networks, including ResNet-18, ResNet-50, Wide ResNet-16-8 (WRN-16-8), VGG-19, and Inception v3. Four adversarial attacks, including Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), Carlini and Wagner (C&W), and Spatial Transformations Attack (ST), were used to complete this investigation. Average accuracy on test images was 96.7% and decreased to 41.1%, 25.5%, 50.1%, and 56.3% in FGSM, PGD, C&W, and ST, respectively. Results are indicating that ResNet-50 and WRN-16-8 were generally less affected by attacks. Therefore using defence methods in these two models can enhance their performance encountering adversarial perturbations.

I. INTRODUCTION

The outbreak of COVID-19 provided new opportunities for machine learning (ML). Applying ML can aid the healthcare system in diagnosing quickly, accurately, and reliably. ML models can diagnose COVID-19 by learning its symptoms. Fever, dry coughs, and lung engagement are general signs of this viral disease [1]. Therefore analyzing lung images along with Reverse Transcription Polymerase Chain Reaction test (RT-PCR) is used as diagnosing methods. Machine Learning (ML) could accelerate the COVID-19 diagnosing procedure while maintaining accuracy and less false negative value by utilizing a wide range of data such as medical images and cough recordings [2], [3].

Convolutional neural networks (CNNs) are specialized architectures in image processing used in medical image classification [4]. IDx-DR is the first U.S. Food and Drug Administration (FDA) approved medical diagnosing system designed to detect diabetic retinopathy [5]. ML-based diagnosing systems were used widely to identify lung engagement from CT-Scan or X-Ray images [6], [7]. According to the capabilities of image classification in rapid diagnosis, several datasets containing lung X-Ray/CT-Scan images were introduced. In addition, Cohen et al. collected confirmed

positive or suspected images of patients with COVID-19 or other viral pneumonia [8].

With all the advancements in Deep Neural Networks (DNNs) robustness to increase precision and accuracy, networks can easily get fooled, putting practical deployment of deep learning systems in jeopardy [9]. Adversarial attack is a procedure that leads to false classification. ML models can be fooled simply even by changing one pixel [10]. Depending on the use of each trained DNN, dreadful results can ensue from attacks. Chernikova et al. studied the effects of such attacks on self-driving cars, in which the prediction of steering angle varies by manipulating input images [11]. In another research, Li et al. studied fooling CNN-based medical diagnosing systems [12]. Hirano et al. investigated universal adversarial perturbation (UAP) on three medical image classification tasks, including skin cancer, referable diabetic retinopathy, and pneumonia images over seven model architectures [13].

Adversarial attacks can be non-targeted or targeted. In a targeted attack, adversarial perturbation affects determined class. Another categorization is based on the access of the attacking kernel to model parameters. Attacking a model by knowing model parameters, gradients, and training process is called a white-box attack. On the contrary, the black-box attack has only access to input and output [14].

In this study, popular DNN models in medical image classification including Residual Networks (RN), Wide Residual Networks (WRN), VGG networks and Inception v3 were investigated.

The rest of the paper organized in three sections. In section II, implemented architectures and attacks are explained. Results reported and discussed in Section III. Finally, we concluded our findings in the last section.

II. MATERIALS AND METHODS

Overall procedure of this investigation is shown in figure 1 and fully described in following subsections.

A. Dataset

Curated Dataset for COVID-19 provides 1281 COVID-19, 3270 Normal, 1656 viral-pneumonia, and 3001 bacterial-pneumonia X-Rays collected from 15 public and available datasets [15]. All images are normalized, standardized, and resized into 255x255 pixels with one color channel, so the final shape is 255x255x1. Additionally, we used normal and COVID-19 sets to train models in binary classification models.

¹Institut National de La Recherche Scientifique (INRS-EMT), University of Quebec Amini.Gougeh.Reza@inrs.ca

Fig. 1: The overall procedure of training and evaluating models

B. DNN Models

Five different DNN architectures with different preferences, depths, and number of parameters were implemented to investigate the effects of adversarial perturbations on their accuracy. RN-18, RN-50, WRN-16-8, VGG-19, and Inception v3 were selected due to their popularity in medical image classification [16], [17]. WRN-L-k is k-level wider than a regular RN with L layers, outperforming common RNs [18]. The number of parameters (trainable and non-trainable) is listed in Table I to compare the complexity of models.

TABLE I: Number of parameters in each model including trainable and non-trainable variables

Models	No. of Parameters
RN-18	11,185,218
RN-50	23,558,978
WRN-16-8	11,033,186
VGG-19	139,577,282
Inception v3	21,806,306

C. Adversarial Attack Kernels

Adversarial Robustness Toolbox (ART) provides a diverse range of attacks, which are used in this study [19].

We focused our investigation on four state-of-art attacks, including Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), Carlini and Wagner (C&W), and Spatial Transformations Attack (ST). The resulted image of each attack may not be noticeable from human eyes, but they can easily fool the DNN model.

FGSM attack uses the gradients of the loss function with respect to the input image to create an adversarial image that maximizes the loss [20].

$$I_{adv} = I + \epsilon \times \text{sign}(\nabla_I L(\theta, I, y)) \quad (1)$$

where I_{adv} is final adversarial image which crafted from I original image. ϵ is to ensure the perturbations are small, ∇_I is gradient, L is the loss function, and θ is model parameters.

On the contrary, the PGD attack is an iterative version of FGSM [21].

$$I_{adv}^t = P_\alpha(I^{t-1} + \epsilon \times \text{sign}(\nabla_I L(\theta, I^{t-1}, y))) \quad (2)$$

Where $P(\cdot)$ is the projection function, and I_{adv}^t is the adversarial image at t-th step. PGD uses random start for $I^0 = I + U_{(-\alpha, \alpha)}$, where U is the uniform distribution.

C&W attack has a different approach to affect models by changing the $L(\cdot)$ function in Eq. (2) [22].

$$I_{adv}^t = P_\alpha(I^{t-1} + \epsilon \times \text{sign}(\nabla_I L(\theta, F(I^{t-1})))) \quad (3)$$

$$F(I^{t-1}) = \max(Z_y(I^{t-1}, \theta) - Z_{y_{target}}(I^{t-1}, \theta), -\tau) \quad (4)$$

where Z is logits regarding y or y_{target} and τ is a parameter to control the confidence of attack.

Despite above mentioned attacks, the ST attack is based on grid search over possible translations and rotations to find optimal attack parameters [23]. The domain of translations for this study was constrained between -30 to 30 degrees

In most of the cases, the final image of implemented attacks (except ST) are not visible to human eyes, Figure 2 illustrates the kernel of attack. Due to the white-box nature of FGSM, PGD, and C&W, the kernel varies with model and hyper-parameters. Also, all implemented attacks are in their default settings in the ART library with infinity norm (L_∞).

Fig. 2: Collection of different types of attacking kernels on implemented models. a) Crafted FGSM kernel on RN-18 b) Crafted PGD kernel on RN-50 c) ST rotated image on WRN-16-8 d) Crafted C&W kernel on VGG-19

III. RESULTS AND DISCUSSION

All of the models were implemented in a Python 3.7 environment from scratch using Kaggle notebooks. Models were trained with 200 epochs, 64 batch sizes, and 10% validation split using Stochastic Gradient Descent (SGD) optimizer. From 4551 images, 3640 were used to train models, and 911 images were used as a test set. While 10% of the train set (364 images) were used in the training validation process.

Table II contains the accuracy of models in classifying original test images and resulted adversarial images by feeding benign test set to FGSM, PGD, C&W, and ST attacking

TABLE II: Accuracy of implemented models over evaluation process using benign and adversarial perturbations

Evaluation set	RN-18	RN-50	WRN-16-8	VGG-19	Inception v3
Test set	93.52	98.35	99.56	97.04	95.39
FGSM	51.9	63.11	38.1	24.84	27.88
PGD	18.48	40.56	30.65	19.06	18.77
C&W	38.63	53.28	56.3	58.71	43.81
ST	56.31	46.65	59.39	65.64	53.9

Fig. 3: The accuracy of each model over testing with five different image sets. Benign test images, FGSM, PGD, C&W and ST.

kernels. Figure 3. Illustrates the accuracy of each model in five different evaluation sets.

The average test accuracy across five DNN models was 96.7%, while this number decreased over testing with adversarial images to the minimum of 26.1% in PGD attack. VGG-19, one of the most commonly used model in medical image classification, was affected by this simple attack, while RN-50 showed moderate resistance. Considering the iterative nature of PGD, the effectiveness of this attack is more than FGSM while C&W attack with its different kernel could reduce the accuracy of the classifiers by more than 40%.

Unlike other attacks, implemented ST attack rotates images from -30 to 30 degrees. The result of this investigation denotes that even slight rotation can reduce the classification accuracy. So training medical image classification models using data augmentation and rotated images is essential.

In overall, PGD attack was more effective than other three attacks. Also, the results demonstrate that WRNs and RN-50 showed more resistance than other models. Therefore securing these models with adversarial retraining can result in a more solid and reliable classification model.

Our findings suggest that selecting a DNN model based on accuracy on benign test set could result in a deceivable system. More evaluations should be considered to check the security, generality, and reliability of the model. According to Table I, number of parameters of VGG-19 could not help this

model to survive attacks, therefore more research is needed to evaluate layer types, activation functions and filter sizes in confronting with adversarial attacks.

IV. CONCLUSIONS

Using ML to classify medical images has been widely investigated by various researchers. One of the main threats against using these models in real applications is adversarial images. In this study, five models were evaluated under four most common attacks. The investigation demonstrated that obtaining acceptable accuracy and Y is not an appropriate criterion for selecting a model out of other architectures and adversarial perturbations could lead to severe problems in such systems. Therefore, adversarial training and model robustness is required for developing practical medical image classification systems. We hope that our findings help researchers enhance their model's security level and increase awareness of developing DNNs with various defense methods.

REFERENCES

- [1] S. Kooraki, M. Hosseiny, L. Myers, and A. Gholamrezanezhad, "Coronavirus (COVID-19) Outbreak: What the Department of Radiology Should Know," *Journal of the American College of Radiology*, vol. 17, no. 4, pp. 447–451, Apr. 2020, doi: 10.1016/j.jacr.2020.02.008.
- [2] D. Dong et al., "The Role of Imaging in the Detection and Management of COVID-19: A Review," *IEEE Rev. Biomed. Eng.*, vol. 14, pp. 16–29, 2021, doi: 10.1109/RBME.2020.2990959.

- [3] J. Laguarda, F. Hueto, and B. Subirana, "COVID-19 Artificial Intelligence Diagnosis Using Only Cough Recordings," *IEEE Open J. Eng. Med. Biol.*, vol. 1, pp. 275–281, 2020, doi: 10.1109/ojemb.2020.3026928.
- [4] S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. K. Khan, "Medical Image Analysis using Convolutional Neural Networks: A Review," *J Med Syst*, vol. 42, no. 11, Oct. 2018, doi: 10.1007/s10916-018-1088-1.
- [5] United States Food and Drug Administration (FDA) News Release. (2018). FDA Permits Marketing of Artificial Intelligence-based Device to Detect Certain Diabetes-related Eye Problems. www.fda.gov/newsevents/newsroom/pressannouncements/ucm604357.htm
- [6] P. Arora, H. Kumar, and B. K. Panigrahi, "Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India," *Chaos, Solitons & Fractals*, vol. 139, p. 110017, Oct. 2020, doi: 10.1016/j.chaos.2020.110017.
- [7] R. Jain, M. Gupta, S. Taneja, and D. J. Hemanth, "Deep learning based detection and analysis of COVID-19 on chest X-ray images," *Appl Intell*, vol. 51, no. 3, pp. 1690–1700, Oct. 2020, doi: 10.1007/s10489-020-01902-1.
- [8] J. P. Cohen, P. Morrison, L. Dao, K., Roth, T. Q., Duong, & M., Ghassemi, "Covid-19 image data collection: Prospective predictions are the future" *arXiv preprint arXiv:2006.11988*, 2020
- [9] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, "Adversarial attacks on medical machine learning," *Science*, vol. 363, no. 6433, pp. 1287–1289, Mar. 2019, doi: 10.1126/science.aaw4399.
- [10] J. Su, D. V. Vargas, and K. Sakurai, "One Pixel Attack for Fooling Deep Neural Networks," *IEEE Trans. Evol. Computat.*, vol. 23, no. 5, pp. 828–841, Oct. 2019, doi: 10.1109/tevc.2019.2890858.
- [11] A. Chernikova, A. Oprea, C. Nita-Rotaru, and B. Kim, "Are Self-Driving Cars Secure? Evasion Attacks Against Deep Neural Networks for Steering Angle Prediction," presented at the 2019 IEEE Security and Privacy Workshops (SPW), May 2019, doi: 10.1109/spw.2019.00033.
- [12] X., Li, D., Pan & D. Zhu, "Defending against adversarial attacks on medical imaging AI system, classification or detection?" *arXiv preprint arXiv:2006.13555*, 2020
- [13] H. Hirano, K. Koga, and K. Takemoto, "Vulnerability of deep neural networks for detecting COVID-19 cases from chest X-ray images to universal adversarial attacks," *PLoS ONE*, vol. 15, no. 12, p. e0243963, Dec. 2020, doi: 10.1371/journal.pone.0243963.
- [14] A., Chakraborty, M., Alam, V., Dey, A., Chattopadhyay & D., Mukhopadhyay, "Adversarial attacks and defences: A survey", *arXiv preprint arXiv:1810.00069*, 2018
- [15] U. Sait, "Curated Dataset for COVID-19 Posterior-Anterior Chest Radiography Images (X-Rays)." *Mendeley*, Sep. 14, 2020, doi: 10.17632/9XKHGTS2S6.1.
- [16] S. Dutta, B. C. S. Manideep, S. Rai, and V. Vijayarajan, "A comparative study of deep learning models for medical image classification," *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 263, p. 042097, Nov. 2017, doi: 10.1088/1757-899x/263/4/042097.
- [17] M. I. Razzak, S. Naz, and A. Zaib, "Deep Learning for Medical Image Processing: Overview, Challenges and the Future," in *Lecture Notes in Computational Vision and Biomechanics*, Springer International Publishing, 2017, pp. 323–350.
- [18] S., Zagoruyko, and N., Komodakis, "Wide residual networks." *arXiv preprint arXiv:1605.07146*, 2016
- [19] M.I., Nicolae, M., Sinn, T.N., Minh, A., Rawat, M., Wistuba, V., Zantedeschi, I.M. Molloy, and B., Edwards., "Adversarial Robustness Toolbox v0. 2.2" 2018.
- [20] I. J., Goodfellow, J., Shlens, & C. Szegedy, "Explaining and harnessing adversarial examples". *arXiv preprint arXiv:1412.6572*.
- [21] A., Madry, A., Makelov, L., Schmidt, D., Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks". *arXiv preprint arXiv:1706.06083*.
- [22] N. Carlini and D. Wagner, "Towards Evaluating the Robustness of Neural Networks," presented at the 2017 IEEE Symposium on Security and Privacy (SP), May 2017, doi: 10.1109/sp.2017.49.
- [23] L., Engstrom, B., Tran, D., Tsipras, L.Schmidt, and A., Madry, "Exploring the landscape of spatial robustness". In *International Conference on Machine Learning* (pp. 1802-1811). PMLR.