Vulnerability of deep neural networks for detecting COVID-19 cases from chest X-ray images to universal adversarial attacks

doi: https://doi.org/10.1371/journal.pone.0243963

PLOS ONE

RESEARCH ARTICLE

Vulnerability of deep neural networks for detecting COVID-19 cases from chest X-ray images to universal adversarial attacks

Hokuto Hirano, Kazuki Koga, Kazuhiro Takemoto*

Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, Iizuka, Fukuoka, Japan

* takemoto@bio.kyutech.ac.jp

OPEN ACCESS

Citation: Hirano H, Koga K, Takemoto K (2020) Vulnerability of deep neural networks for detecting COVID-19 cases from chest X-ray images to universal adversarial attacks. PLoS ONE 15(12): e0243963. https://doi.org/10.1371/journal.pone.0243963

Editor: Haoran Xie, Lingnan University, HONG KONG

Received: July 13, 2020

Accepted: December 2, 2020

Published: December 17, 2020

Data Availability Statement: The code used in this study is available from our GitHub repository: github.com/hkthirano/UAP-COVID-Net. The chest X-ray images used in this study are publicly available online (see github.com/lindawangg/COVID-Net/blob/master/docs/COVIDx.md for details).

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Owing the epidemic of the novel coronavirus disease 2019 (COVID-19), chest X-ray computed tomography imaging is being used for effectively screening COVID-19 patients. The development of computer-aided systems based on deep neural networks (DNNs) has become an advanced open source to rapidly and accurately detect COVID-19 cases because the need for expert radiologists, who are limited in number, forms a bottleneck for screening. However, thus far, the vulnerability of DNN-based systems has been poorly evaluated, although realistic and high-risk attacks using universal adversarial perturbation (UAP), a single (input image agnostic) perturbation that can induce DNN failure in most classification tasks, are available. Thus, we focus on representative DNN models for detecting COVID-19 cases from chest X-ray images and evaluate their vulnerability to UAPs. We consider non-targeted UAPs, which cause a task failure, resulting in an input being assigned an incorrect label, and targeted UAPs, which cause the DNN to classify an input into a specific class. The results demonstrate that the models are vulnerable to non-targeted and targeted UAPs, even in the case of small UAPs. In particular, the 2% norm of the UAPs to the average norm of an image in the image dataset achieves >85% and >90% success rates for the non-targeted and targeted attacks, respectively. Owing to the non-targeted UAPs, the DNN models judge most chest X-ray images as COVID-19 cases. The targeted UAPs allow the DNN models to classify most chest X-ray images into a specified target class. The results indicate that careful consideration is required in practical applications of DNNs to COVID-19 diagnosis; in particular, they emphasize the need for strategies to address security concerns. As an example, we show that iterative fine-tuning of DNN models using UAPs improves the robustness of DNN models against UAPs.

Introduction

Coronavirus disease 2019 (COVID-19) [1] is an infectious disease caused by the coronavirus, called severe acute respiratory syndrome coronavirus 2. The COVID-19 epidemic started from Wuhan, China [2], and has had a severe impact on public health and the economy globally [3]. To reduce the spread of this epidemic, effective screening of COVID-19 patients is required.

Thus, positive real-time polymerase chain reaction (PCR) tests are mainly used [4]; however, they are often time consuming and laborious and involve complicated manual processes. Chest radiography, especially chest X-ray computed tomography (CT) imaging, becomes an alternative screening method [5] because patients present abnormalities in chest radiography images, which are a characteristic of those infected with COVID-19 [2, 6]. Moreover, there are advantages to leveraging chest X-ray imaging for COVID-19 screening amid the pandemic in terms of rapid triaging, portability, availability, and accessibility [7]. However, the visual differences in chest X-ray images among COVID-19-associated pneumonia, non-COVID-19 pneumonia, and no pneumonia are subtle; thus, the need for expert radiologists, who are limited in number, forms a bottleneck for diagnoses based on radiography images. To overcome this limitation, computer-aided systems that can aid radiologists in more rapidly and accurately interpreting radiography images to detect COVID-19 cases are highly required [7, 8]; in particular, deep neural networks (DNNs) are often used for this purpose.

DNNs are widely used for image classification, a task in which an input image is assigned a class from a fixed set of classes as well as medical science [9, 10], including diagnoses based on radiography images. Specifically, DNN-based systems can detect subtle visual differences in the images; in particular, a DNN can accurately distinguish bacterial and viral pneumonia in chest X-ray images [11]. Inspired by these previous studies, many researchers have constructed large-scale datasets of chest radiography images on COVID-19 [7, 8, 12, 13] and have proposed DNN-based systems for screening COVID-19 cases from these images [8, 14–17]. However, DNN-based systems in medical science have generally been closed source and unavailable to the research community for deeper understanding and extension. Thus, Wang et al. [7] proposed COVID-Net, a deep convolutional neural network design intended to detect COVID-19 cases from chest X-ray images. COVID-Net is one of the first open-source network designs for COVID-19 detection. As the authors mentioned [7], this study will be leveraged and built upon by both researchers and citizen data scientists to accelerate the development of highly accurate yet practical deep learning solutions for detecting COVID-19 cases and accelerate the treatment of the disease. The COVID-Net models are intended to be used as reference models; in fact, several DNN-based systems [18–20] for detecting COVID-19 cases have already been proposed, inspired by the COVID-Net study.

However, previous studies have poorly evaluated the vulnerabilities in DNNs, although DNNs are known to be vulnerable to adversarial examples [21, 22], which are input images that cause misclassifications by DNNs and are usually generated by adding specific, imperceptible perturbations to original input images that have been correctly classified using DNNs. Adversaries can easily attack open-sourced software, such as COVID-Net because they can access the model parameters and training data; thus, it is important to evaluate the reliability and safety of DNNs against adversarial attacks.

These adversarial attacks may be less useful for adversaries because they are input image dependent (i.e., an individual adversarial perturbation is used such that each input image is misclassified). However, more realistic adversarial attacks have been proposed in recent years. Notably, a single perturbation (called *universal adversarial perturbation*, *UAP*, as they are image agnostic) [23] that can induce DNN failure in most image classification tasks also exists. UAPs are difficult to detect because such perturbations are extremely small and, hence, do not significantly affect data distributions. UAP-based adversarial attacks can be more straightforward to implement by adversaries in real-world environments. A previous study [23] considered only UAPs for non-targeted attacks, which cause misclassification (i.e., a task failure resulting in an input image being assigned an incorrect class). However, we previously extended the algorithm for generating UAPs to enable targeted attacks [24], causing the DNN to classify an input image into a specific class. The existence of adversarial examples questions

the generalization ability of DNNs, reduces model interpretability, and limits the applications of deep learning in safety- and security-critical environments [25]. Specifically, vulnerability is a severe problem in medical diagnosis [26]. Thus, it is important to evaluate the vulnerability of the proposed DNN-based systems to adversarial attacks (attacks based on UAPs, in particular) in practical applications. In addition, defense strategies against adversarial attacks (i.e., adversarial defense [22]) are required.

In this study, we focus on the COVID-Net models, which are representative models for detecting COVID-19 cases from chest X-ray images, and aim to evaluate the vulnerability of DNNs to adversarial attacks. Specifically, the vulnerability to non-targeted and targeted attacks, based on UAPs, is investigated. Moreover, adversarial defense is considered; in particular, we evaluate to what extent the robustness of COVID-Net models to non-targeted and targeted UAPs increases using adversarial retraining [23, 27] (i.e., fine-tuning with adversarial images).

Material and methods

COVID-Net models

We forked the COVID-Net repository (github.com/lindawangg/COVID-Net) on May 1, 2020, and obtained two DNN models for detecting COVID-19 cases from chest X-ray images: COVIDNet-CXR Small and COVIDNet-CXR Large. Moreover, we downloaded the COVIDx dataset, a collection of chest radiography images from several open-source chest radiography datasets, on May 1, 2020, according to the description in the COVID-Net repository. The chest X-ray images in the dataset were classified into three classes: *normal* (no pneumonia), *pneumonia* (non-COVID-19 pneumonia; e.g., viral and bacterial pneumonia), and *COVID-19* (COVID-19 viral pneumonia). The dataset comprised 13,569 training images (7,966 *normal* images, 5,451 *pneumonia* images, and 152 *COVID-19* images) and 231 test images (100 *normal* images, 100 *pneumonia* images, and 31 *COVID-19* images).

Universal adversarial perturbations

The UAPs for non-targeted and targeted attacks were generated using simple iterative algorithms [23, 28], whose details are described in [23, 28]. We used the non-targeted UAP algorithm available in the Adversarial Robustness 360 Toolbox (ART) [29] (version 1.0; github.com/IBM/adversarial-robustness-toolbox). The targeted UAP algorithm was implemented by modifying the non-targeted UAP algorithm in the ART in our previous study [24] (github.com/hkthirano/targeted_UAP_CIFAR10).

The algorithms consider a classifier, $C(\boldsymbol{x})$, which returns the class or label with the highest confidence score for an input image, $\boldsymbol{x}$. The algorithm starts with $\boldsymbol{\rho} = \mathbf{0}$ (no perturbation) and iteratively updates the UAP, $\boldsymbol{\rho}$, under the constraint that the L_p norm of the perturbation is equal to or less than a small ξ value (i.e., $\|\boldsymbol{\rho}\|_p \leq \xi$), by additively obtaining an adversarial perturbation for an input image, $\boldsymbol{x}$, which is randomly selected from an input image set, $\boldsymbol{X}$, without replacement. These iterative updates continue until the number of iterations reaches a maximum $i_{\max}$.

We used the fast gradient sign method (FGSM) [21] to obtain an adversarial perturbation for the input image, instead of the original UAP algorithm [23], which uses the DeepFool method [30]. This is because FGSM is used for both non-targeted and targeted attacks, and DeepFool requires a higher computational cost than FGSM and only generates a non-targeted adversarial example for the input image. FGSM generates the adversarial perturbation, $\hat{\boldsymbol{\rho}}$, for $\boldsymbol{x}$ using gradient $\nabla_x L(\boldsymbol{x}, y)$ of the loss function at the specified image $\boldsymbol{x}$ and class y with respect to the pixels [21]. For the L_∞ norm, a non-targeted perturbation that causes misclassification is

computed as $\hat{\boldsymbol{\rho}} = \epsilon \cdot \text{sign}(\nabla_x L(\boldsymbol{x}, C(\boldsymbol{x})))$, whereas a targeted perturbation that causes C classification of an image $\boldsymbol{x}$ into class y is obtained as $\hat{\boldsymbol{\rho}} = -\epsilon \cdot \text{sign}(\nabla_x L(\boldsymbol{x}, y))$, where ϵ (> 0) is the attack strength. For the L_1 and L_2 norms, a non-targeted perturbation is computed as $\hat{\boldsymbol{\rho}} = \epsilon \cdot \nabla_x L(\boldsymbol{x}, C(\boldsymbol{x}))/ \parallel \nabla_x L(\boldsymbol{x}, C(\boldsymbol{x}))_p \parallel$, whereas a targeted perturbation is obtained as $\hat{\boldsymbol{\rho}} = -\epsilon \cdot \nabla_x L(\boldsymbol{x}, y)/\parallel \nabla_x L(\boldsymbol{x}, y) \parallel_p$.

In the algorithms, FGSM is performed based on the output $C(\boldsymbol{x} + \boldsymbol{\rho})$ of the classifier for the perturbed image $\boldsymbol{x} + \boldsymbol{\rho}$, at each iteration step. For non-targeted (targeted) attacks, an adversarial perturbation, $\hat{\boldsymbol{\rho}}$, for $\boldsymbol{x} + \boldsymbol{\rho}$ is obtained using the FGSM if $C(\boldsymbol{x} + \boldsymbol{\rho}) = C(\boldsymbol{x}) \cdot (C(\boldsymbol{x} + \boldsymbol{\rho}) \neq y)$. After generating the adversarial example (i.e., $\boldsymbol{x}_{\text{adv}} \leftarrow \boldsymbol{x} + \boldsymbol{\rho} + \hat{\boldsymbol{\rho}}$) at this step, the perturbation $\boldsymbol{\rho}$ is updated if $C(\boldsymbol{x}_{\text{adv}}) \neq C(\boldsymbol{x})$ ($C(\boldsymbol{x}_{\text{adv}}) = y$) for non-targeted (targeted) attacks. When updating $\boldsymbol{\rho}$, a projection function project, $(\boldsymbol{x}, p, \xi)$, is used to satisfy the constraint that $\|\boldsymbol{\rho}\|_p \leq \xi$: $\boldsymbol{\rho} \leftarrow$ project$(\boldsymbol{x}_{\text{adv}} - \boldsymbol{x}, p, \xi)$, where project$(\boldsymbol{x}, p, \xi) = \arg\min_{x'} \|\boldsymbol{x} - \boldsymbol{x}'\|_2$ subject to $\|\boldsymbol{\rho}\|_p \leq \xi$.

The non-targeted and targeted UAPs were generated using 13,569 training images in the COVIDx dataset. Parameter ϵ was set to 0.001; the cases where $p = 2$ and ∞ were considered. Meanwhile, parameter ξ was determined based on the ratio ζ of the L_p norm of the UAP to the average L_p norm of an image in the COVIDx dataset. Cases in which $\zeta = 1\%$ and 2% (i.e., almost imperceptible perpetuations) were considered. The average L_∞ and L_2 norms were 237 and 32,589, respectively; i_{max} was set to 15.

To compare the performance of the generated UAPs with that of random controls, we also generated random vectors (random UAPs) sampled uniformly from the sphere of a specified radius [23].

Vulnerability evaluation

To evaluate the vulnerability of the DNN models to UAPs, we used the fooling rate, R_f, and targeted the attack success rate, R_s, of non-targeted and targeted attacks, respectively. The R_f of an image set is defined as the proportion of images that were not classified into their associated actual labels to all images in the set. The R_s of an image set is the proportion of adversarial images classified into the target class to all images in the set. Additionally, we obtained the confusion matrices to evaluate the change in prediction owing to the UAPs for each class (infection type).

Adversarial retraining

We performed adversarial retraining to increase the robustness of the COVID-Net models to UAPs [23, 27]; in particular, the models were fine-tuned with adversarial images, and the procedure was described in a previous study [23]. A brief description is provided below. 1) Ten UAPs against a DNN model were generated using the algorithm (for generating a non-targeted or targeted UAP) (see Materials and methods section) with the (clean) training image set. 2) A modified training image set was obtained by randomly selecting half of the training images and combining them with the rest, where each image was perturbed by a UAP randomly selected from 10 UAPs. 3) The model was fine-tuned by performing five extra epochs of training on the modified training image set. 4) A new UAP (against the fine-tuned model) was generated using the algorithm with the training image set. 5) R_f and R_s of the UAP for the test images were then computed. Steps 1)–5) were repeated five times.

Results

Performance of COVID-Net models

The test accuracies of the COVIDNet-CXR Small and COVIDNet-CXR Large models were 92.6% and 94.4%, respectively, and their training accuracies were 95.8% and 94.1%,

respectively. As shown in the COVID-Net study [7], we also confirmed that the COVID-Net models achieved good accuracies.

Vulnerability to non-targeted universal adversarial perturbations

However, we found that both COVIDNet-CXR Small and COVIDNet-CXR Large models were vulnerable to non-targeted UAPs (Table 1). Specifically, the fooling rate, R_f, of the UAPs with $\zeta = 1\%$ for the test image set was 81.0% at most. A higher ζ led to a higher R_f. We observed that the R_f of the UAP with $\zeta = 2\%$ for the test image set was between 85.7% and 87.4%. Furthermore, the random UAPs with $\zeta = 2\%$ misclassified the models; specifically, their R_f were up to 22.1%. The change in R_f did not exhibit significant dependence on the norm types ($p = 2$ or ∞). The difference in R_f for the test image set between $p = 2$ and $p = \infty$ was up to 7%, the model and the other parameters being equal. R_f of the UAP against the COVIDNet-CXR Small model was lower than that of the COVIDNet-CXR Large model in the case of $\zeta = 1\%$, the model and the other parameters being equal; however, no remarkable difference in R_f between these models was observed in the case of $\zeta = 2\%$. The R_f of the training image set was higher than that of the test image set because the UAPs were generated based on the training image set.

Owing to non-targeted UAPs, the models classified most images into *COVID-19*. Fig 1 shows the confusion matrices for the COVID-Net models attacked using non-targeted UAPs with $p = \infty$. For the UAPs with $\zeta = 1\%$, the COVIDNet-CXR Small model classified >70% of the *normal* and *pneumonia* test images into *COVID-19*. Moreover, the COVIDNet-CXR Large model classified approximately 90% of the *normal* and *pneumonia* images into *COVID-19*. For a higher ζ, this tendency was more significant. In particular, the COVIDNet-CXR Small and Large models evaluated almost all *normal* and *pneumonia* test images as COVID-19 cases when $\zeta = 2\%$. Additionally, the tendency of adversarial images to be classified into *COVID-19* was observed when considering UAPs with $p = 2$ and the training image set.

The non-targeted UAPs with $\zeta = 1\%$ and $\zeta = 2\%$ were almost imperceptible. Fig 2 shows the non-targeted UAPs $p = \infty$ against the COVID-Net models and their adversarial images. The models classified the original X-ray images (left panels in Fig 2) and correctly predicted their actual classes; however, they evaluated all adversarial images as COVID-19 cases owing to the non-targeted UAPs. Similarly, the non-targeted UAPs $p = 2$ were almost imperceptible.

Vulnerability to targeted universal adversarial perturbations

Furthermore, we found that both the COVIDNet-CXR Small model (Table 2) and COVID-Net-CXR Large model (Table 3) were vulnerable to targeted UAPs. Subsequently, we considered the effect of the targeted attacks using UAPs in each class: *normal*, *pneumonia*, and

Table 1. Fooling rates R_f (%) of non-targeted UAPs against the COVID-Net models.

p	ζ	COVIDNet-CXR Small		COVIDNet-CXR Large	
		Training	Test	Training	Test
2	1%	61.4 (1.3)	58.0 (0.4)	90.0 (2.5)	81.0 (3.9)
	2%	98.5 (12.6)	87.4 (16.0)	97.4 (17.9)	85.7 (22.1)
∞	1%	70.8 (1.0)	64.9 (1.3)	84.8 (2.0)	77.1 (3.5)
	2%	98.5 (9.4)	87.4 (13.4)	97.4 (14.3)	85.7 (19.9)

The R_f of the training and test images are presented. The values in the brackets indicate R_f random UAPs (random controls).

https://doi.org/10.1371/journal.pone.0243963.t001

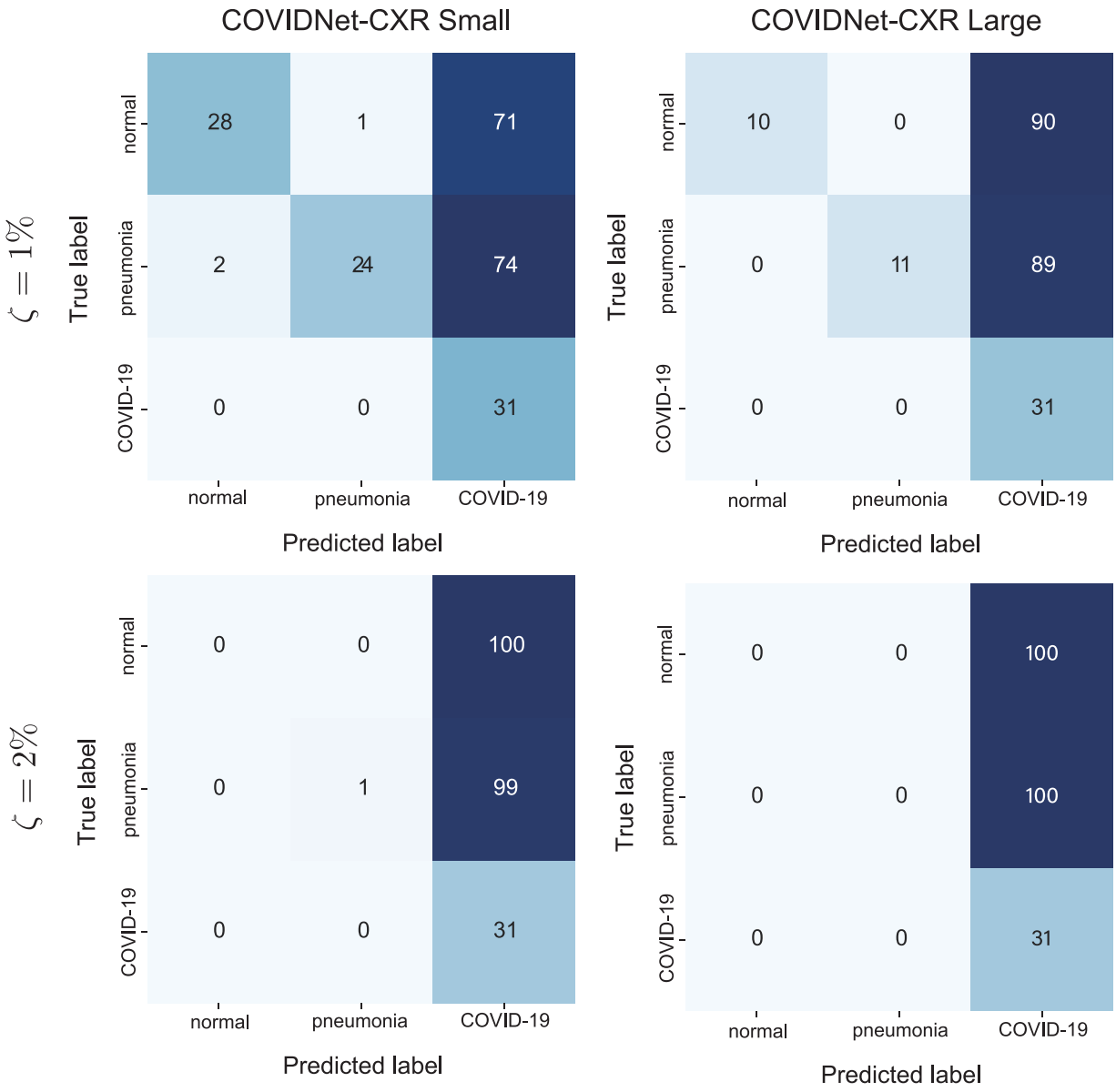

Fig 1. Confusion matrices for the COVID-Net models attacked using the non-targeted UAPs on the test images. $p = \infty$. Left and right panels represent the COVIDNet-CXR Small and COVIDNet-CXR Large models, respectively. The top and bottom panels indicate $\zeta = 1\%$ and $\zeta = 2\%$, respectively.

https://doi.org/10.1371/journal.pone.0243963.g001

COVID-19. When $\zeta = 1\%$, the targeted attack success rates R_s for the test images were between approximately 60% and 85% and between approximately 55% and 95% for the COVIDNet-CXR Small and Large models, respectively. Conversely, the R_s of the training images was between approximately 65% and 90% and between approximately 55% and 90%. Meanwhile, the R_s of the UAP with $p = 2$ was higher than that of the UAP with $p = \infty$, the model, and the other parameters being equal. Moreover, no remarkable difference in the R_s was observed between the target classes; however, the R_s of the target attacks to *COVID-19* were relatively high in the COVIDNet-CXR Large model. Thus, a higher ζ led to a higher R_s. When $\zeta = 2\%$, the R_s values for both the training and test images were approximately 100%, regardless of the

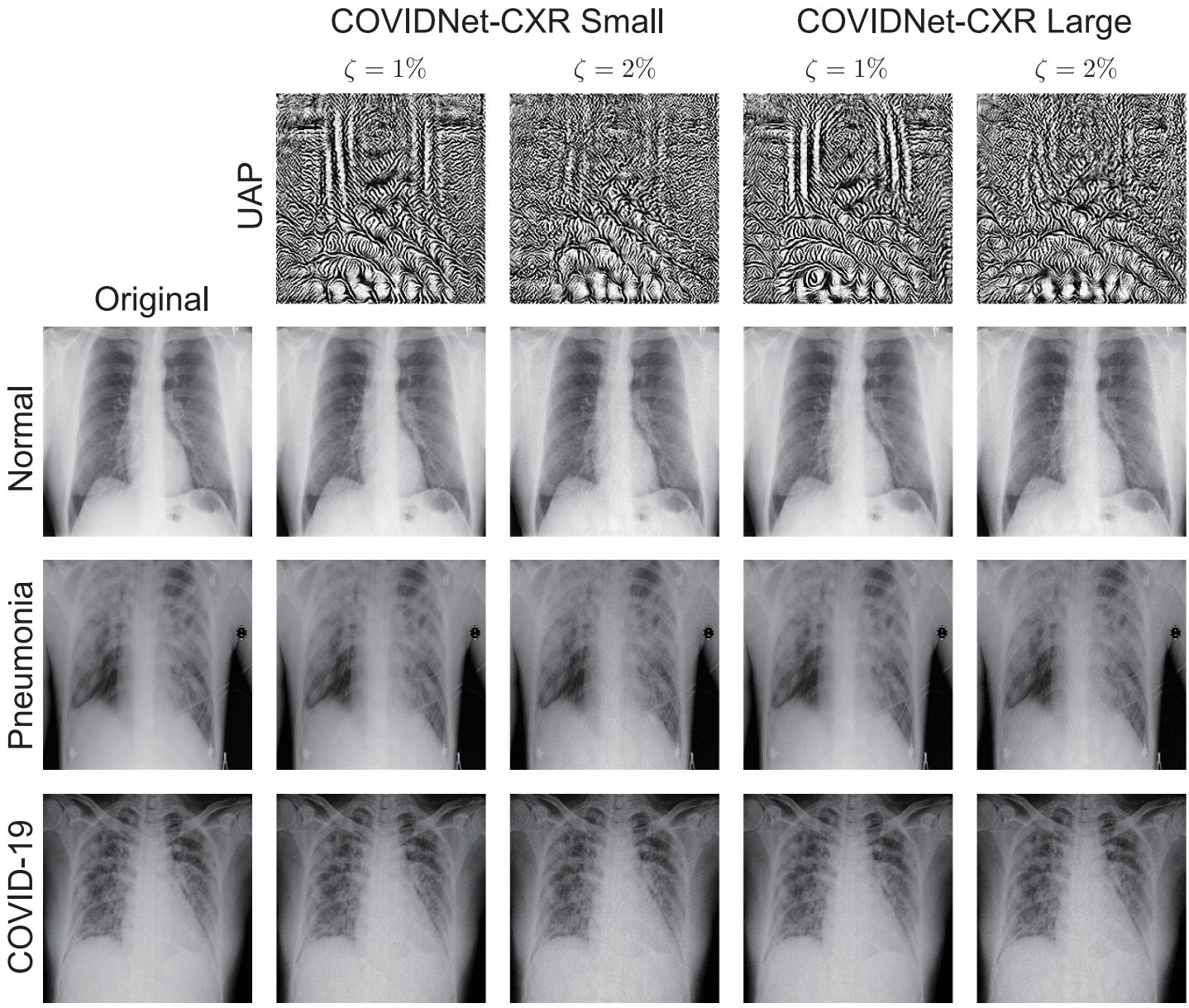

Fig 2. Non-targeted UAPs with $p = \infty$ against the COVID-Net models and their adversarial images. UAPs (top panels) with ζ = 1% and ζ = 2% are shown. The models correctly classified the original images (left panels) into their actual labels. The predicted labels of all adversarial images are of *COVID-19*. Note that the UAPs are emphatically displayed for clarity; in particular, each UAP is scaled by a maximum of 1 and a minimum of 0.

https://doi.org/10.1371/journal.pone.0243963.g002

target classes. For the targeted attacks to *normal* and *pneumonia*, the R_s of random UAPs for the test images were also relatively high; in particular, they were between approximately 35% and 45% and between approximately 30% and 45% for the COVIDNet-CXR Small model and COVIDNet-CXR Large model, respectively.

It was difficult to classify the *COVID-19* images into another targeted class (*normal* or *pneumonia*) when the UAPs were relatively weak (i.e., ζ = 1%). Fig 3 shows the confusion matrices for the COVIDNet-CXR Small model attacked using targeted UAPs with $p = \infty$. For both targeted attacks to *normal* and *pneumonia*, the model correctly predicted almost all *COVID-19* images as COVID-19 cases, despite the targeted attacks. Conversely, approximately 50% of

Table 2. Targeted attack success rate R_s (%) of targeted UAPs against the COVIDNet-CXR Small model to each target class.

p	ζ	*Normal*		*Pneumonia*		*COVID-19*	
		Training	Test	Training	Test	Training	Test
2	1%	88.1 (60.5)	78.4 (46.3)	76.7 (37.5)	71.4 (41.6)	68.1 (1.9)	74.0 (12.1)
	2%	99.4 (54.4)	97.8 (39.0)	99.4 (33.0)	98.7 (35.9)	100 (12.6)	99.1 (25.1)
∞	1%	79.5 (60.7)	64.9 (45.9)	66.5 (37.5)	61.9 (41.6)	78.8 (1.8)	84.0 (12.6)
	2%	98.7 (56.3)	96.1 (39.4)	99.5 (34.1)	98.3 (37.7)	100 (9.5)	100 (22.9)

The R_s for the training and test images are shown in Table 2. The values in brackets are R_s random UAPs (random controls).

https://doi.org/10.1371/journal.pone.0243963.t002

normal (*pneumonia*) images were classified as targeted class *pneumonia* (*normal*). However, for a higher ζ (i.e., $\zeta = 2\%$), the targeted attacks of the *COVID-19* images were successful; in particular, almost all *COVID-19* images were classified into the target class (*normal* or *pneumonia*) because of the UAP. The classification of the images into COVID-19 using targeted UAPs was easier than that into the other classes. Owing to the UAP with $\zeta = 1\%$, the model judged approximately 80% of *normal* and *pneumonia* images as COVID-19 cases, respectively. Similar tendencies were observed in the COVIDNet-CXR Large model for targeted UAPs with $p = 2$ and on the training image set.

The targeted UAPs were also almost imperceptible. Fig 4 shows the targeted UAPs with $p = \infty$ and $\zeta = 2\%$ against the COVIDNet-CXR Small model and their adversarial images. The model classified the original images (left panels in Fig 4) and correctly predicted their actual classes (source classes); however, it classified the adversarial images into each target class because of the targeted UAPs. The UAPs with $\zeta = 1\%$ were also imperceptible. Additionally, imperceptibility was confirmed in the UAPs with $p = 2$ and those against the COVIDNet-CXR Large model.

Effect of adversarial retraining

Adversarial retraining is often used to avoid adversarial attacks. In this study, we investigated the extent to which adversarial retraining increases the robustness of the COVIDNet-CXR Small model to non-targeted and targeted UAPs with $p = \infty$. Adversarial retraining did not affect the test accuracy in either non-targeted or targeted cases; specifically, the accuracy on the (clean) test images remained constant at approximately 90% (Fig 5A and 5B).

For non-targeted attacks using UAPs with $\zeta = 2\%$, R_f for the test images declined with the iterations for adversarial retraining; in particular, it was 22.1% after five iterations (Fig 5A). The confusion matrix (Fig 5C) for the fine-tuned model obtained after five iterations indicates that the *normal* and *COVID-19* images were almost correctly classified despite the non-targeted UAPs. However, 45% of the *pneumonia* images were still misclassified.

Table 3. Targeted attack success rates R_s (%) of targeted UAPs against the COVIDNet-CXR Large model to each target class.

p	ζ	*Normal*		*Pneumonia*		*COVID-19*	
		Training	Test	Training	Test	Training	Test
2	1%	85.2 (58.9)	71.4 (44.2)	72.6 (37.0)	66.2 (39.0)	92.4 (4.0)	95.2 (16.9)
	2%	99.2 (50.7)	98.3 (34.6)	99.5 (30.6)	98.7 (32.9)	100 (18.7)	100 (32.5)
∞	1%	71.0 (59.2)	56.7 (44.2)	55.4 (37.0)	53.2 (40.3)	88.4 (3.7)	92.2 (15.6)
	2%	97.9 (52.7)	93.9 (35.9)	99.4 (32.3)	98.3 (33.8)	100 (14.9)	100 (30.3)

The R_s for the training and test images are shown in Table 3. The values in brackets are R_s random UAPs (random controls).

https://doi.org/10.1371/journal.pone.0243963.t003

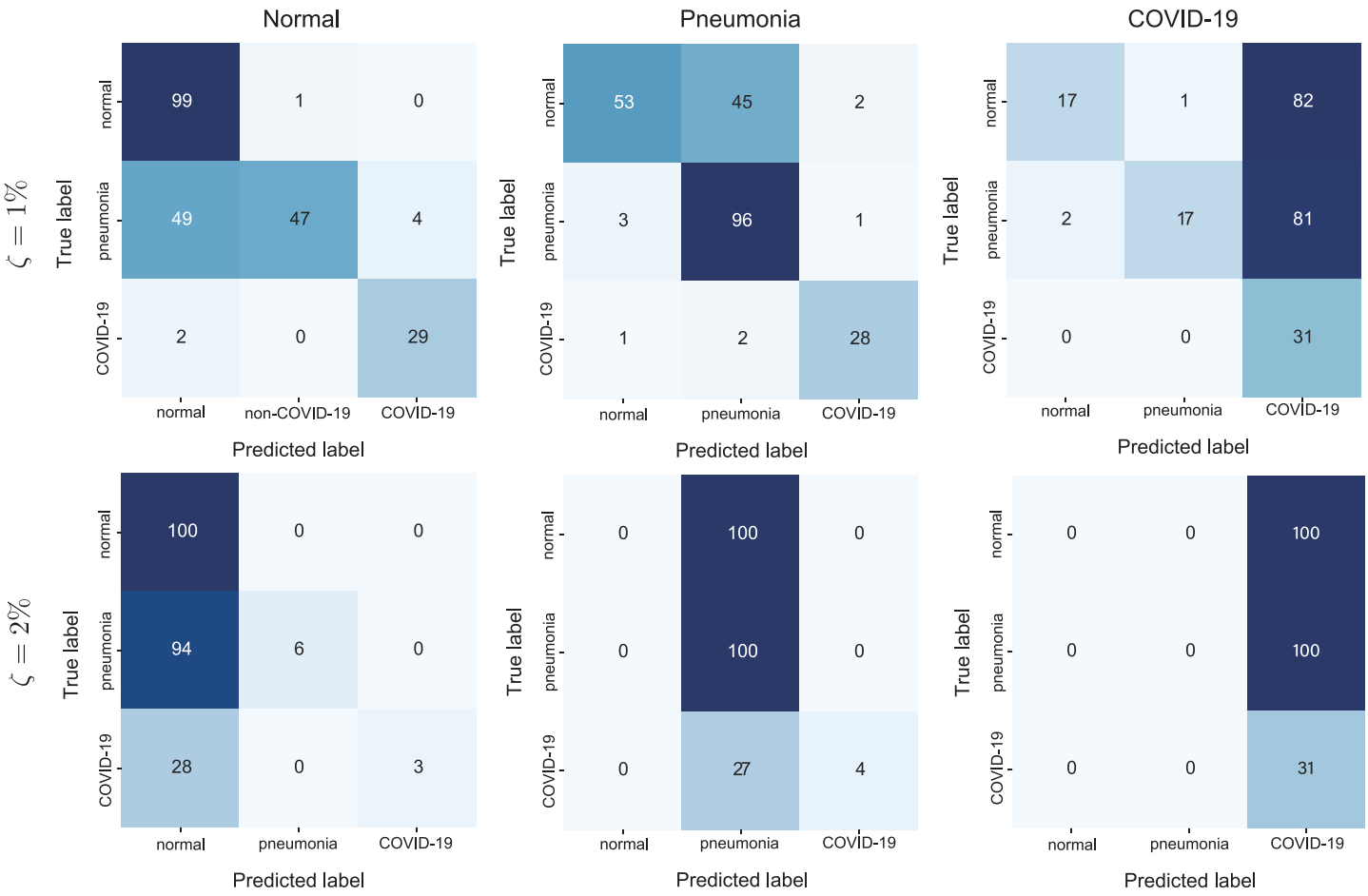

Fig 3. Confusion matrices for the COVIDNet-CXR Small model attacked with the targeted UAPs with $p = \infty$ on the test images. The left, middle, and right panels represent the targeted classes: *normal*, *pneumonia*, and *COVID-19*, respectively. The top and bottom panels indicate $\zeta = 1\%$ and $\zeta = 2\%$, respectively.

https://doi.org/10.1371/journal.pone.0243963.g003

For targeted attacks to *COVID-19* using UAPs with $\zeta = 1\%$, the R_s for the test images decreased with the iterations for adversarial retraining (Fig 5B); specifically, it was 16.5% after five iterations. The confusion matrix (Fig 5D) for the fine-tuned model obtained after five iterations indicates that the *normal* and *COVID-19* images were almost correctly classified despite the targeted UAPs. However, 15% of the *pneumonia* images were still misclassified as *COVID-19*.

Discussion

The COVID-Net models were vulnerable to small UAPs; moreover, they were slightly less robust to random UAPs. The results indicated that the DNN-based systems were easy to mislead. Adversaries can result in failing the DNN-based systems at lower costs (i.e., using a single perturbation); specifically, they do not need to consider the distribution and diversity of input images when attacking the DNNs using UAPs, as UPAs are image agnostic. Considering that vulnerability to UAPs is observed in various DNN architectures [23, 24], they are expected to exist universally in DNN-based systems for detecting COVID-19 cases.

For non-targeted attacks with UAPs, the COVID-Net models predicted most of the chest X-ray images as COVID-19 cases because of the UAPs (Fig 1), although the UAPs were almost

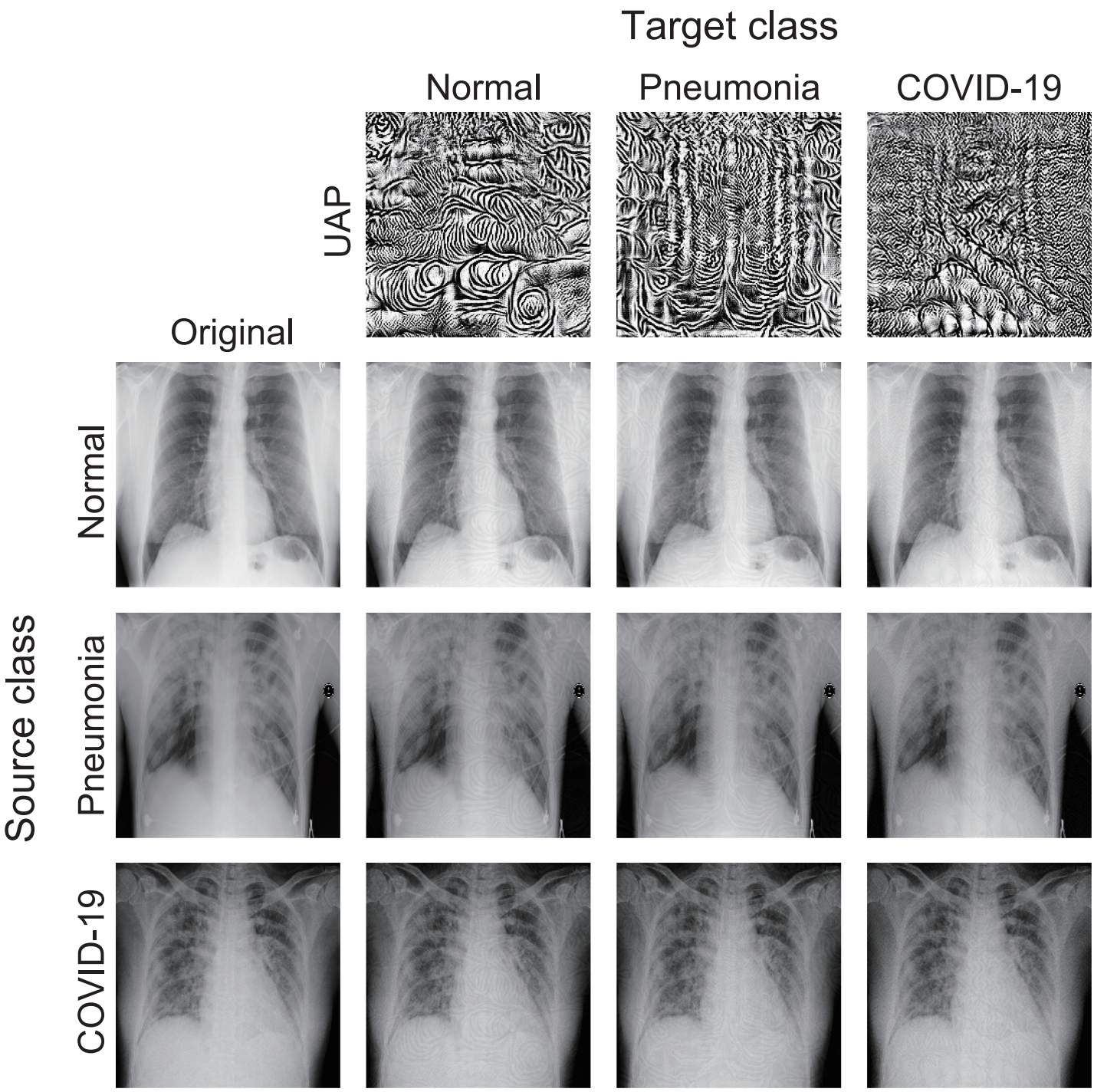

Fig 4. Targeted UAPs (top panel) with $\zeta = 2\%$ and $p = \infty$ against the COVIDNet-CXR Small model and their adversarial images. Note that UAPs are emphatically displayed for clarity; in particular, each UAP is scaled by a maximum of 1 and a minimum of 0.

https://doi.org/10.1371/journal.pone.0243963.g004

imperceptible (Fig 2). This result is consistent with the tendency of DNN models to classify most inputs into a few specific classes because of non-targeted UAPs (i.e., existence of dominant labels in non-targeted attacks based on UAPs) [23]. Moreover, this indicates that the models provide false positives in COVID-19 diagnosis, which may cause unwanted mental

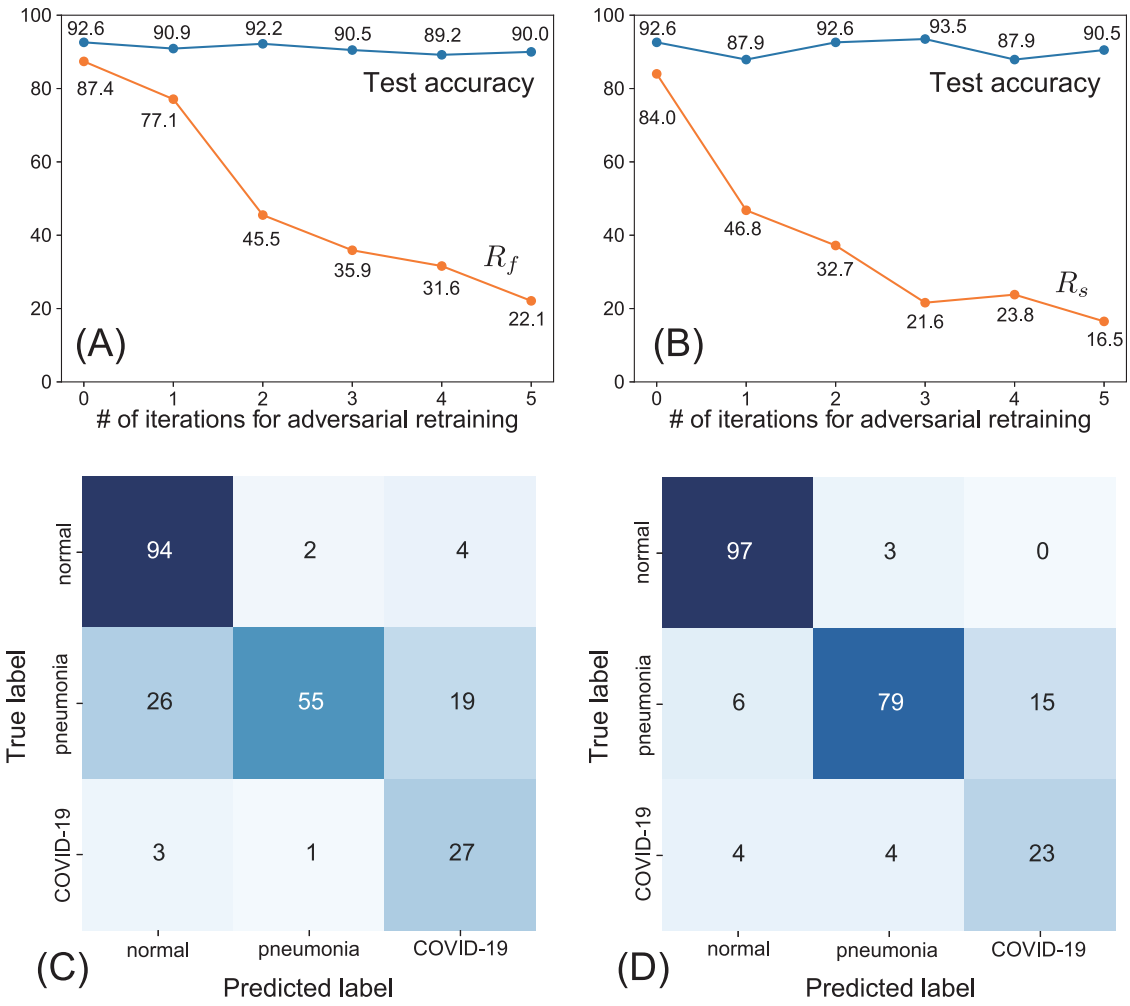

Fig 5. Effect of adversarial retraining on the robustness to UAPs with $p = \infty$ against the COVIDNet-CXR Small model. Scatter plots of (A) the fooling rate, R_f (%), for non-targeted UAPs with $\zeta = 2\%$ versus the number, N_i, of iterations for adversarial retraining and (B) the targeted attack success rate, R_s (%), of targeted UAPs with $\zeta = 1\%$ to *COVID-19* versus N_i. Here, R_f and R_s are for the test images. The accuracies (%) on the set of clean test images are also shown. The confusion matrices for the fine-tuned models were obtained after five iterations of adversarial retraining using the (C) non-targeted UAPs and (D) targeted UAPs. Note that these confusion matrices belong to the fine-tuned models attacked using non-targeted and targeted UAPs, respectively.

https://doi.org/10.1371/journal.pone.0243963.g005

stress to patients and complicate the estimation of the number of COVID-19 cases. The dominant label of COVID-19 observed in this study may be because the COVIDx dataset was imbalanced. The images in *COVID-19* were predominantly fewer than those in *normal* and *pneumonia* cases. The algorithm considers maximizing the fooling rate; thus, a relatively large fooling rate is achieved when all inputs are classified into *COVID-19* because of UAPs. In addition, the observed dominant label may be because the losses were computed by weighting the *COVID-19* class to consider the imbalanced dataset. The decision for the *COVID-19* class might be more susceptible to changes in pixel values than that for the other classes.

The relatively easy targeted attacks on *COVID-19* (Fig 3) may be because *COVID-19* was the dominant label. Moreover, targeted attacks to *normal* and *pneumonia* were possible, despite almost imperceptible UAPs (Fig 4). The results imply that adversaries can control DNN-based systems, which may lead to security concerns. The targeted attacks cause both false positives and negatives, and thus, can be used to adjust the number of COVID-19 cases.

Moreover, they may affect individual and social awareness of COVID-19 (e.g., voluntary restraint and social distancing). These may lead to problems in terms of public health (i.e., minimizing the spread of the pandemic) and the economy. More generally, complex classifiers, including DNNs, are currently used for high-stake decision making in healthcare; however, they can potentially cause catastrophic harm to the society because they are often difficult to interpret [31].

The COVID-Net models, with tailored network architecture, seem to be more vulnerable to adversarial attacks than representative DNN models (e.g., VGG [32] and ResNet [33] models) for classifying ideal natural images (e.g., CIFAR-10 [34] and ImageNet datasets [35]). For these representative DNNs, UAPs with $\zeta = 5\%$ and higher are required to achieve $>80\%$ success rates for non-targeted and targeted attacks [23, 28]. Conversely, for the COVID-Net models, UAPs with $\zeta = 2\%$ achieved $>85\%$ and $>90\%$ success rates for the non-targeted and targeted attacks, respectively. This result implies several possible reasons that caused the vulnerability of COVID-Net models. For example, the variance (visual difference) in chest X-ray images is much less than that in natural images. In this case, data points may aggregate around decision boundaries, indicating that the outputs of the DNN models are susceptible to changes in pixel values. As a result, adversarial examples are easy to generate. In addition, the fact that adversarial vulnerability of DNNs is known to increase with input dimension [36] may be one of the causes.

The UAPs used in this study are a type of white-box attack, which assumes that adversaries can access the model parameters (the gradient of the loss function, in this case) and training images; thus, they are security threats for open-source software projects, such as COVID-Net. A simple solution to prevent these adversarial attacks is to make DNN-based systems closed-source and publicly unavailable; however, this conflicts with the purpose of accelerating the development of computer-based systems for detecting COVID-19 cases and COVID-19 treatment. An alternative may be to consider black-box systems, such as closed application programming interfaces (APIs) and closed-source software in which only queries on inputs are allowed and outputs are accessible. Such closed APIs are better because they are at least publicly available. However, it is possible that APIs are vulnerable to adversarial attacks. This is because UAPs have generalizability [23] (i.e., UAPs for a DNN can mislead another DNN). That is, adversarial attacks on black-box DNN-based systems may be possible using the UAPs generated based on white-box DNNs. Moreover, several methods for adversarial attacks on black-box DNN-based systems, which estimate adversarial perturbations using only model outputs (e.g., confidence scores), have been proposed [37–39].

Therefore, defense strategies against adversarial attacks should be considered. A simple defense strategy is to fine-tune DNN models using adversarial images [22, 23, 27]. In fact, we demonstrated that iterative fine-tuning of a DNN model using UAPs improved the robustness of the DNN model to non-targeted and targeted UAPs (Fig 5). However, the iterative fine-tuning method required high computational costs, and it did not perfectly avoid vulnerability to UAPs. In addition, several methods breaching defenses using adversarial retraining have already been proposed [27]. Alternatively, dimensionality reduction (e.g., principle component analysis), distributional detection (e.g., maximum mean discrepancy), and normalization detection (e.g., dropout randomization) may be useful for adversarial defenses; however, adversarial examples are not easily detected using these approaches [27]. Defending against adversarial attacks is a cat-and-mouse game [26]; thus, it may be difficult to completely avoid security concerns caused by adversarial attacks. However, the development of methods for defending against adversarial attacks has advanced. For example, detecting adversarial attack-based robustness to random noise [40], the use of a discontinuous activation function that purposely invalidates the DNN's gradient at densely distributed input data points [41], and DNNs for purifying adversarial examples [42] may help reduce the concerns.

In conclusion, we demonstrated the vulnerability of DNNs for detecting COVID-19 cases to non-targeted and targeted attacks based on UAPs. However, many studies have developed DNN-based systems for detecting COVID-19 while ignoring the vulnerability. Our findings emphasize that careful consideration is required in developing DNN-based systems for detecting COVID-19 cases and their practical applications. Facile applications of DNNs to COVID-19 detection could lead to problems in terms of public health and the economy. Our study is the first to show the vulnerability of DNNs for COVID-19 detection and to alert such facile applications of DNNs. The code used in this study is available from our GitHub repository: github.com/hkthirano/UAP-COVID-Net. The chest X-ray images used in this study are publicly available online (see github.com/lindawangg/COVID-Net/blob/master/docs/COVIDx.md for details).

Acknowledgments

The authors are much obliged to Dr. Seyed-Mohsen Moosavi-Dezfooli for his helpful comments regarding the fine-tuning of DNN models with UAPs. The authors would like to thank Editage (www.editage.com) for English language editing.

Author Contributions

Conceptualization: Kazuhiro Takemoto.

Data curation: Hokuto Hirano.

Formal analysis: Hokuto Hirano, Kazuhiro Takemoto.

Investigation: Hokuto Hirano, Kazuki Koga, Kazuhiro Takemoto.

Methodology: Hokuto Hirano, Kazuki Koga, Kazuhiro Takemoto.

Project administration: Kazuhiro Takemoto.

Software: Hokuto Hirano.

Supervision: Kazuhiro Takemoto.

Validation: Hokuto Hirano, Kazuhiro Takemoto.

Visualization: Hokuto Hirano, Kazuhiro Takemoto.

Writing – original draft: Hokuto Hirano, Kazuhiro Takemoto.

Writing – review & editing: Kazuhiro Takemoto.

References

1. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis. 2020; https://doi.org/10.1016/S1473-3099(20)30120-1 PMID: 32087114
2. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet. 2020; 395: 497–506. https://doi.org/10.1016/S0140-6736(20)30183-5 PMID: 31986264
3. Ahmed F, Ahmed N, Pissarides C, Stiglitz J. Why inequality could spread COVID-19. Lancet Public Heal. 2020; https://doi.org/10.1016/S2468-2667(20)30085-2 PMID: 32247329
4. Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China. JAMA. 2020; 323: 1061. https://doi.org/10.1001/jama.2020.1585 PMID: 32031570
5. Fang Y, Zhang H, Xie J, Lin M, Ying L, Pang P, et al. Sensitivity of chest CT for COVID-19: comparison to RT-PCR. Radiology. 2020; 200432. https://doi.org/10.1148/radiol.2020200432 PMID: 32073353

6. Ng M-Y, Lee EY, Yang J, Yang F, Li X, Wang H, et al. Imaging profile of the COVID-19 infection: radiologic findings and literature review. Radiol Cardiothorac Imaging. 2020; 2: e200034. https://doi.org/10.1148/ryct.2020200034
7. Wang L, Wong A. COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-Ray images. 2020; http://arxiv.org/abs/2003.09871
8. Zhang K, Liu X, Shen J, Li Z, Sang Y, Wu X, et al. Clinically applicable AI system for accurate diagnosis, quantitative measurements and prognosis of COVID-19 pneumonia using computed tomography. Cell. 2020; https://doi.org/10.1016/j.cell.2020.04.045 PMID: 32416069
9. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal. Elsevier B.V.; 2017; 42: 60–88. https://doi.org/10.1016/j.media.2017.07.005 PMID: 28778026
10. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Heal. The Author(s). Published by Elsevier Ltd. This is an Open Access article under the CC BY 4.0 license; 2019; 1: e271–e297. https://doi.org/10.1016/S2589-7500(19)30123-2
11. Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell. Elsevier Inc.; 2018; 172: 1122–1131.e9. https://doi.org/10.1016/j.cell.2018.02.010 PMID: 29474911
12. Zhao J, Zhang Y, He X, Xie P. COVID-CT-Dataset: a CT scan dataset about COVID-19. 2020; 2003.13865
13. Cohen JP, Morrison P, Dao L. COVID-19 image data collection. 2020; 2003.11597
14. Zhang J, Xie Y, Li Y, Shen C, Xia Y. COVID-19 screening on chest X-ray images using deep learning based anomaly detection. 2020; http://arxiv.org/abs/2003.12338
15. Wang L, Lin ZQ, Wong A. COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. Sci Rep. 2020; 10: 19549. https://doi.org/10.1038/s41598-020-76550-z PMID: 33177550
16. Tartaglione E, Barbano CA, Berzovini C, Calandri M, Grangetto M. Unveiling COVID-19 from chest X-ray with deep learning: a hurdles race with small data. 2020; http://arxiv.org/abs/2004.05405
17. Lv D, Qi W, Li Y, Sun L, Wang Y. A cascade network for detecting COVID-19 using chest X-rays. 2020; http://arxiv.org/abs/2005.01468
18. Farooq M, Hafeez A. COVID-ResNet: a deep learning framework for screening of COVID19 from radiographs. 2020; http://arxiv.org/abs/2003.14395
19. Afshar P, Heidarian S, Naderkhani F, Oikonomou A, Plataniotis KN, Mohammadi A. COVID-CAPS: a capsule network-based framework for identification of COVID-19 cases from X-ray Images. 2020; http://arxiv.org/abs/2004.02696
20. Rahimzadeh M, Attar A. A new modified deep convolutional neural network for detecting COVID-19 from X-ray images. 2020; http://arxiv.org/abs/2004.08052
21. Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. 2014; http://arxiv.org/abs/1412.6572
22. Yuan X, He P, Zhu Q, Li X. Adversarial examples: attacks and defenses for deep learning. IEEE Trans Neural Networks Learn Syst. 2019; 30: 2805–2824. https://doi.org/10.1109/TNNLS.2018.2886017 PMID: 30640631
23. Moosavi-Dezfooli SM, Fawzi A, Fawzi O, Frossard P. Universal adversarial perturbations. Proc—30th IEEE Conf Comput Vis Pattern Recognition, CVPR 2017. 2017;2017-Janua: 86–94. 10.1109/CVPR.2017.17
24. Hirano H, Takemoto K. Simple iterative method for generating targeted universal adversarial perturbations. Proceedings of 25th International Symposium on Artificial Life and Robotics. 2020. pp. 426–430. http://arxiv.org/abs/1911.06502
25. Matyasko A, Chau L-P. Improved network robustness with adversary critic. 2018; http://arxiv.org/abs/1810.12576
26. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. Science (80-). 2019; 363: 1287–1289. https://doi.org/10.1126/science.aaw4399 PMID: 30898923
27. Carlini N, Wagner D. Adversarial examples are not easily detected. Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security—AISec '17. New York, New York, USA: ACM Press; 2017. pp. 3–14. 10.1145/3128572.3140444

28. Hirano H, Takemoto K. Simple iterative method for generating targeted universal adversarial perturbations. Algorithms. 2020; 13: 268. https://doi.org/10.3390/a13110268

29. Nicolae M-I, Sinn M, Tran MN, Buesser B, Rawat A, Wistuba M, et al. Adversarial Robustness Toolbox v1.0.0. 2018; http://arxiv.org/abs/1807.01069

30. Moosavi-Dezfooli S-M, Fawzi A, Frossard P. DeepFool: a simple and accurate method to fool deep neural networks. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2016. pp. 2574–2582. 10.1109/CVPR.2016.282

31. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019; 1: 206–215. https://doi.org/10.1038/s42256-019-0048-x

32. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 3rd International Conference on Learning Representations, ICLR 2015—Conference Track Proceedings. 2015.

33. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2016. pp. 770–778. 10.1109/CVPR.2016.90

34. Krizhevsky A. Learning Multiple Layers of Features from Tiny Images. Tech report, Univ Toronto. 2009; 10.1.1.222.9220

35. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet Large Scale Visual Recognition Challenge. Int J Comput Vis. 2015; https://doi.org/10.1007/s11263-015-0816-y

36. Simon-Gabriel C-J, Ollivier Y, Bottou L, Schölkopf B, Lopez-Paz D. First-order adversarial vulnerability of neural networks and input dimension. Proceedings of the 36th International Conference on Machine Learning (ICML). PMLR; 2019. pp. 5809–5817. http://proceedings.mlr.press/v97/simon-gabriel19a.html

37. Chen J, Su M, Shen S, Xiong H, Zheng H. POBA-GA: Perturbation optimized black-box adversarial attacks via genetic algorithm. Comput Secur. 2019; 85: 89–106. https://doi.org/10.1016/j.cose.2019.04.014

38. Guo C, Gardner JR, You Y, Wilson AG, Weinberger KQ. Simple black-box adversarial attacks. Proc 36th Int Conf Mach Learn. 2019; 2484–2493. http://arxiv.org/abs/1905.07121

39. Co KT, Muñoz-González L, de Maupeou S, Lupu EC. Procedural noise adversarial examples for black-box attacks on deep convolutional networks. Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. New York, NY, USA: ACM; 2019. pp. 275–289. 10.1145/3319535.3345660

40. Yu T, Hu S, Guo C, Chao W-L, Weinberger KQ. A new defense against adversarial images: turning a weakness into a strength. Adv Neural Inf Process Syst. 2019; 1633–1644.: 1910.07629

41. Xiao C, Zhong P, Zheng C. Enhancing adversarial defense by k-winners-take-all. Proc 8th Int Conf Learn Represent. 2020; http://arxiv.org/abs/1905.10510

42. Hwang U, Park J, Jang H, Yoon S, Cho NI. PuVAE: a variational autoencoder to purify adversarial examples. IEEE Access. 2019; 7: 126582–126593. https://doi.org/10.1109/ACCESS.2019.2939352