Natural Images Allow Universal Adversarial Attacks on Medical Image Classification Using Deep Neural Networks With Transfer Learning

Akinori Minagi  
Kyushu Institute of Technology

Hokuto Hirano  
Kyushu Institute of Technology

Kazuhiro Takemoto  (✉ takemoto@bio.kyutech.ac.jp)  
Kyushu Institute of Technology

Research Article

Keywords: deep neural networks, transfer learning, medical imaging, adversarial attacks, security and privacy

Posted Date: October 1st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-757225/v2

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

Version of Record: A version of this preprint was published at Journal of Imaging on February 4th, 2022. See the published version at https://doi.org/10.3390/jimaging8020038.
Natural images allow universal adversarial attacks on medical image classification using deep neural networks with transfer learning

Akinori Minagi¹, Hokuto Hirano¹, Kazuhiro Takemoto¹*
¹) Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, Iizuka, Fukuoka 820-8502, Japan
*Corresponding author’s e-mail: takemoto@bio.kyutech.ac.jp

Abstract

Background. Transfer learning from natural images is well used in deep neural networks (DNNs) for medical image classification to achieve computer-aided clinical diagnosis. Although the adversarial vulnerability of DNNs hinders practical applications owing to the high stakes of diagnosis, adversarial attacks are expected to be limited because training data — which are often required for adversarial attacks — are generally unavailable in terms of security and privacy preservation. Nevertheless, we hypothesized that adversarial attacks are also possible using natural images because pre-trained models do not change significantly after fine-tuning.

Methods. We considered three representative DNN-based medical image classification tasks (i.e., skin cancer, referable diabetic retinopathy, and pneumonia classifications) to investigate whether medical DNN models with transfer learning are vulnerable to universal adversarial perturbations (UAPs), generated using natural images.

Results. UAPs from natural images are useful for both non-targeted and targeted attacks. The performance of UAPs from natural images was significantly higher than that of random controls, although slightly lower than that of UAPs from training images. Vulnerability to UAPs from natural images was observed between different natural image datasets and between different model architectures.

Conclusion. The use of transfer learning causes a security hole, which decreases the reliability and safety of computer-based disease diagnosis. Model training from random initialization (without transfer learning) reduced the performance of UAPs from natural images; however, it did not completely avoid vulnerability to UAPs. The vulnerability of UAPs from natural images will become a remarkable security threat.

Keywords: deep neural networks, transfer learning, medical imaging, adversarial attacks, security and privacy
Background

Transfer learning from natural image datasets (e.g., the ImageNet dataset [1]) is a widely used technique in deep neural networks (DNNs) for image classification; in particular, it has been well applied to medical imaging [2]. Although the amount of medical image data is often limited, transfer learning enables the acquisition of highly accurate DNNs from such limited image data by fine-tuning existing model architectures (e.g., Inception V3 [3] and ResNet50 [4]), pre-trained using the ImageNet dataset with the image data. The transfer learning technique has been used for skin cancer classification based on photographic images [5], referable diabetic retinopathy classification based on optical coherence tomography (OCT) images of the retina [6], and pneumonia classification based on chest X-ray images [6]. The diagnostic performance of these DNNs is high and equivalent to that of healthcare professionals [7]. Thus, DNNs with transfer learning are beginning to be applied to medical image diagnosis to empower physicians and accelerate decision-making in clinical environments [2].

However, the practical application of DNNs to disease diagnosis may still be debatable owing to the existence of adversarial examples [8–10]; these are input images that are typically generated by adding specific, imperceptible perturbations to the original input images, leading to DNN misclassification. Given that diagnosing disease involves making high-stake decisions, the existence of adversarial examples is a security concern [11]. Adversarial examples probably cause misdiagnosis and various social disturbances [12], as well as limiting deep learning applications in safety- and security-critical environments [13]. Therefore, it is also important to evaluate the reliability and safety of DNNs against adversarial attacks in medical imaging.

Many previous studies have demonstrated that DNN models are vulnerable to input-dependent adversarial attacks (i.e., an individual adversarial perturbation is used such that each input image is misclassified) in skin cancer [12] and pneumonia classifications [14]. More importantly, a previous study [15] showed that a single small image agnostic perturbation, called universal adversarial perturbation (UAP) [16, 17], can induce DNN failure in most tasks and is also a security threat to DNN-based medical image classifications. UAP-based attacks are more realistic because they can be more easily implemented by adversaries in real-world environments and are practicable, with a lower computational cost [16].

A simple solution to avoid adversarial attacks is to render training data and any other similar domain-specific data (e.g., medical images in the case of medical image classification) publicly unavailable because various methods of adversarial attacks [8–10] (from attack methods that assume access to DNN model weights to those that do not) generally assume the use of such data to generate adversarial perturbations. Given that the data availability of medical images is generally limited in terms of security and privacy preservation [11], adversarial attacks on DNN-based medical image classifications seem to be limited. However, we doubt this prediction because of the properties of transfer learning for medical imaging [18]. Specifically, transfer learning considers that model weights pre-trained with the ImageNet dataset (natural images) are fine-tuned with medical images; however, fine-tuned DNN models for medical imaging are known to be almost similar to
the original pre-trained DNN models, despite the fine-tuning process. Additionally, larger DNN models do not change through training. It seems that DNN models obtained by fine-tuning well-used model architectures (e.g., Inception V3 and ResNet50) with medical images show almost similar reactions to both medical and natural images.

Thus, we developed and tested the hypothesis that adversarial perturbations against fine-tuned DNN models are generatable using not only training data (medical images) but also natural images (e.g., the ImageNet dataset). Following our previous study [15], we considered representative medical image classifications (skin cancer classification based on photographic images [5], referable diabetic retinopathy classification based on OCT images [6], and pneumonia classification based on chest X-ray images [6]) and investigated the vulnerability of fine-tuned DNN models with several architectures to adversarial perturbations generated using natural images. In this study, we focused on universal adversarial attacks [16, 17] rather than input-dependent adversarial attacks. This is because input-dependent adversarial attacks are less effective; in particular, it is costly to determine the medical images that result in misclassification from an adversarial perturbation generated using a natural image. In contrast, UAPs (generated using natural images) can be used for any medical image because they are image agnostic. To evaluate the effects of transfer learning on vulnerability to UAPs, we also considered the DNN model architecture training from random initialization.

**Methods**

**Medical image datasets and models**

We used the medical image datasets and DNN models previously described in [15] (see also github.com/hkthirano/MedicalAI-UAP). A brief description is provided below:

The skin lesion images for skin cancer classification comprised 7,000 training images and 3,015 test images that were classified into seven classes: melanoma (MEL), melanocytic nevus (NV), basal cell carcinoma (BCC), actinic keratosis/Bowens disease (intraepithelial carcinoma; AKIEC), benign keratosis (solar lentigo/seborrheic keratosis/lichen planus-like keratosis; BKL), dermatofibroma (DF), and vascular lesions (VASC). The OCT images for referable diabetic retinopathy classification consisted of 7,840 training and 3,360 test images that were classified into four classes: choroidal neovascularization with neovascular membrane and associated subretinal fluid (CNV), diabetic macular edema with retinal-thickening-associated intraretinal fluid (DME), multiple drusen present in early age-related macular degeneration (DRUSEN), and normal retina with preserved foveal contour and absence of any retinal fluid/edema (NM). The chest X-ray images for pneumonia classification comprised 1,800 training and 540 test images that were classified into binary classes: no pneumonia (NORMAL) or viral or bacterial pneumonia (PNEUMONIA). Note that the OCT and chest X-ray image datasets were class-balanced but the skin lesion image dataset was not.

The Inception V3 architecture [3] was mainly considered, following previous studies [5, 6]. To evaluate the effect of model architecture on vulnerability to UAPs, we also used the VGG16 [19] and ResNet50 [4] architectures. These DNN model architectures pre-trained using the ImageNet dataset were fine-tuned with the training images in a medical image
dataset, using the learning rate schedule and data augmentation. We also trained the
Inception V3 models with the training images in a medical image dataset from random
initialization to evaluate the effects of transfer learning on the vulnerability to UAPs from
natural images. The training conditions in this case (e.g., the learning rate schedule,
condition of data augmentation, and number of epochs) were identical to those in the case
of transfer learning, except for random initialization.

Universal adversarial perturbations and natural images

Following our previous study[15], simple iterative algorithms [16, 17] were used to
generate UAPs. We considered both non-targeted attacks, which cause misclassification
(i.e., a task failure resulting in an input image being assigned an incorrect class), and
targeted attacks, which caused the DNN to classify an input image into a specific class. for
the non-targeted and targeted attacks, respectively. The Adversarial Robustness 360
Toolbox (ART) [20] (version 1.0; github.com/Trusted-AI/adversarial-robustness-toolbox)
was used. For targeted UAPs, we used our proposed method [17] (see also
github.com/hkthirano/targeted_UAP_CIFAR10), which was implemented by modifying
the nontargeted UAP algorithm [16].

The algorithms consider a classifier and generate UPAs $\rho$ from an input image set $X$,
under the constraint that the $L_p$ norm of the perturbation $||\rho||_p \leq \xi$, for a small $\xi$ value.
The algorithms start with $\rho = 0$ (no perturbation) and iteratively update $\rho$ by additively
obtaining an adversarial perturbation for an input image $x$, which is randomly selected
from $X$ without replacement via the fast gradient sign method [8] with the attack strength
parameter $\epsilon$. These iterative updates continue until the number of iterations reaches a
maximum $i_{\text{max}}$.

Using these algorithms, UAPs against a DNN model for medical image classification were
generated using natural images. We considered the training images in the ImageNet dataset
because the DNN models were pre-trained using the ImageNet dataset. The ImageNet
training set was downloaded from www.image-net.org/download.php on June 17, 2020.
Moreover, we also considered the Open Images dataset (V6), a different dataset of natural
images, to evaluate the dataset dependency on the performance of UAPs. The dataset was
downloaded from storage.googleapis.com/openimages/web/download.html on November
22, 2020. For each dataset, 100,000 randomly selected images were used to generate the
UAPs. The images were gray-transformed when generating UAPs against the DNN models
for referable diabetic retinopathy and pneumonia classifications.

For skin lesion and chest X-ray image classifications, the parameters $\epsilon$ and $p$ were set to
0.0005 and 2, respectively. For the OCT image classification, $\epsilon$ and $p$ were set to 0.0013
and $\infty$, respectively. However, different $\epsilon$ was considered for the Inception V3 models
trained from random initialization. When generating UAPs using training images, $\epsilon$ was
0.0044, 0.0036, and 0.0066 for the skin lesion, OCT, and chest X-ray image classifications,
respectively. When generating UAPs using natural images, $\epsilon$ was 0.0050, 0.0020, and
0.0026 for the skin lesion, OCT, and chest X-ray image classifications, respectively. The
parameters $\epsilon$ and $p$ were selected using a grid search to maximize the performance of
the UAPs (see below) for input images. The parameter $i_{\text{max}}$ was set to 1. The parameter
ξ was set based on the ratio ζ of the $L_p$ norm of the UAP to the average $L_p$ norm of an image in the dataset (see [15] for the actual values of the average $L_p$ norms).

We also obtained the UAPs generated using the training images in a medical dataset from our previous study [15] to compare the performances of the UAPs between the training images and natural images. Random vectors (random UAPs) were sampled uniformly from the sphere of a specified radius to compare the performances of the generated UAPs with those of the random controls [16].

**Evaluating the performance of UAPs**

Evaluating the performance of UAPs was based on the procedures established in our previous study [15]. The fooling rate $R_f$ and targeted attack success rate $R_s$ was used to evaluate the performance of a nontargeted UAP ($\rho_{nt}$) and targeted UAP ($\rho_t$). $R_f = |X|^{-1} \sum_{x \in X} \mathbb{I}(C(x) \neq C(x + \rho_{nt}))$, where $C(x)$ be an output (class or label) of a classifier (DNN) for an input image $x$ in an image set $X$. The function $\mathbb{I}(A)$ returns 1 if the condition $A$ is true and 0 otherwise. $R_f$ indicates the fraction of adversarial images from which the labels predicted are inconsistent with the labels predicted from clean images to all images in the set. $R_s = |X|^{-1} \sum_{x \in X} \mathbb{I}(C(x + \rho_t) = y)$: the proportion of adversarial images classified into the target class $y$ to all images in set $X$. Note that $R_s$ has a baseline, $R_s$ observed without UAPs. As mentioned in our previous study [15], the label composition of the image data and prediction performance of DNNs both affect the baseline. The $R_s$ baselines of UAPs targeted to a specified class were ~25% and ~50%, respectively, for the OCT and chest X-ray image datasets. For the skin lesion dataset, the $R_s$ baselines of UAPs targeted to MEL and NV were ~10% and ~65%, respectively. $R_f$ and $R_s$ were computed using test images from the medical image dataset. The confusion matrices on test images from the medical image dataset were also obtained to evaluate the change in prediction owing to the UAPs for each class. The rows and columns in the matrices represent the true and predicted classes, respectively. The confusion matrices were row-normalized to account for imbalanced datasets.

**Results**

**Natural images allow nontargeted universal adversarial attacks on medical image classification**

We first consider the Inception V3 models as it was used by previous studies on DNN-based medical imaging [5, 6] and evaluated whether nontargeted UAPs against the DNN models for medical image classification are generatable using natural images (Fig. 1). The performance of the UAPs generated using the natural images was less effective than that of the UAPs generated in the training images in the medical image dataset; specifically, the UAPs from the training images achieved a higher fooling rate $R_f$, with a smaller perturbation magnitude $\zeta$, compared to the UAPs from the natural images. However, $R_f$ of the UAPs generated using the natural images was significantly higher than that of random UAPs; moreover, they also increased rapidly with $\zeta$ and reached a high $R_f$, despite a low $\zeta$. Specifically, $R_f$ ~80% and ~50% were achieved at $\zeta = 4\%$ for the skin...
lesion (Fig. 1a) and chest X-ray image classifications (Fig. 1c), respectively. \( R_f \) was 40–60\% at \( \zeta = 8\% \) for the OCT image classification (Fig. 1b). These UAPs were almost imperceptible. As a representative example, clean images and their adversarial examples owing to the UAPs from the ImageNet dataset are shown in Fig. 2. These results indicate that small UAPs from natural images also cause misclassification of DNN-based medical image classifications. We also found that the performance of UAPs from natural images has no strong dataset dependency because \( R_f \) of the UAPs from the Open Images dataset was almost similar to those of the UAPs generated using the ImageNet dataset, although small differences in \( R_f \) were observed, i.e., ~40\% and ~60\% for the Open images and ImageNet datasets, respectively.

For the ResNet50 and VGG16 models, \( R_f \) of the UAPs from the natural images was also significantly higher than that of the random control (Fig. 3), although it was less than that of the UAPs from the training images. However, \( R_f \) at the same \( \zeta \) was different between the model architectures, except for the chest X-ray image classification. For the skin lesion image classification (Fig. 3a), \( R_f \) of the UAPs with \( \zeta = 4\% \) was approximately 80\% for the Inception V3 model, whereas it was lower for the ResNet50 and VGG16 models. Specifically, \( R_f \) against the ResNet50 model and the VGG model were ~70\% and 30–50\%, respectively. For the OCT image classification (Fig. 3b), a slightly higher \( R_f \) (60–70\%) of the UAPs against ResNet50 and VGG16 models with \( \zeta = 8\% \) was observed, compared to the Inception V3 model (40–60\%). For the chest X-ray image classification (Fig. 3c), \( R_f \) of the UAPs with \( \zeta = 4\% \) from the natural images was ~50\%, independent of the model architecture.

As expected from the observed difference in \( R_f \) between the UAPs from training images, and those generated from natural images, those from the natural images were visually different to those from the training images, for the same \( \zeta \). Figure 4 shows the UAPs generated using the training, ImageNet, and Open Images datasets against Inception V3 models. Moreover, the confusion matrices on the test images also showed a different tendency of misclassification of the DNN models due to UAPs between the UAPs from the natural images and those from the training images, although the confusion matrix patterns are similar in that the DNN models classify most images into several specific classes (i.e., dominant classes) due to UAPs. Figure 5 shows the Inception V3 model. For the skin lesion image classification, the dominant classes were MEL and BLK when using the UAPs from the training images; however, the dominant class was only MEL when using the UAPs from the natural images (both the ImageNet and Open Images datasets). For the OCT image classification, the dominant class was CNV in the case of the UAPs from the training images; however, it was DRUSEN and NM in the case of the UAPs from the ImageNet dataset and in the case of the UAPs from the Open Images dataset. For the chest X-ray image classification, the DNN model incorrectly predicted the true labels because of the UAPs from the training images; however, it classified most images into NORMAL because of the UAPs from the natural images (both the ImageNet and Open Images datasets), indicating that \( R_f \) saturated at ~50\% (Fig. 1c).

The dominant classes might differ according to the model architectures and natural image datasets, except for the chest X-ray image classification. For the skin lesion classification,
the dominant class was BKL for the UAPs from both the ImageNet and Open Images datasets against the VGG16 model and for the UAP from the Open Images dataset against ResNet50, whereas it was MEL for the UAPs from the ImageNet dataset against the ResNet50 model (Additional file 1: Fig. S1). For the OCT image classification, the dominant classes of the UAPs from both the ImageNet and Open Images datasets were DRUSEN for ResNet50; however, they were CNVs for the VGG16 model (Additional file 1: Fig. S2). For the chest X-ray image classification, the dominant classes were NORMAL, independent of the model architectures and natural image datasets (Additional file 1: Fig. S3).

**Natural images allow targeted universal adversarial attacks on medical image classification**

We also investigated whether the targeted UAPs were generatable using natural images. Following our previous study [15], we considered targeted attacks to be the most significant case and the control in each medical image dataset. The most significant cases correspond to MEL, CNV, and PNEUMONIA in the skin lesion, OCT, and chest X-ray image datasets, respectively. The controls correspond to NV, NM, and NORMAL in the skin lesion, OCT, and chest X-ray image datasets, respectively. Table 1 presents the targeted attack success rate $R_s$ of the UAPs against DNN-based medical image classification. The UAPs were very small and almost imperceptible because $\zeta = 4\%$ in the skin lesion and chest X-ray image classifications and $\zeta = 8\%$ in the OCT image classification, as in the case of non-targeted UAPs (see Fig. 2). However, overall, the values of $R_s$ (>90%) of the UAPs from both the ImageNet and Open Images datasets were significantly higher than those of random UAPs; moreover, they were almost similar to those of the UAPs from the training images in the medical image dataset. This tendency was independent of the model architectures. However, low $R_s$ was observed in a small number of cases. $R_s$ of the UAPs from the ImageNet and Open Images targeted to MEL were ~10%, which were almost similar to the random control for the ResNet50 model, whereas they were ~95% for the Inception V3 and ResNet50 models. $R_s$ of the UAPs from the ImageNet and Open Images targeted to CNV were 35 – 50 %, which were higher than random controls for the ResNet50 model, whereas they were ~100% for Inception V3 and VGG16 models. $R_s$ of the UAPs from the ImageNet and Open Images targeted to PNEUMONIA were 60 – 80%, which was higher than that of random controls, whereas that of the UAPs from the training images was ~100%.

As representative examples, Figure 6 shows the UAPs generated using several image datasets for targeted attacks on MEL, CNV, and NORMAL against Inception V3 models. These UAPs showed $R_s$ ~100%; however, the UAPs from natural images were visually different from those from the training images for each medical image dataset.

**Effect of transfer learning on the vulnerability of the UAPs from natural images**

It is predicted that transfer learning from natural images (the ImageNet dataset, in particular) causes the observed vulnerability of the UAPs from natural images to DNN-based medical image classification. To test this more deeply, we considered the Inception V3 models, which are widely used in medical image classification [5, 6], trained with the
training images in each medical image dataset from random initialization. The test
accuracies of the models were 79.2%, 95.3%, and 97.8% for the skin lesion, OCT, and
chest X-ray image datasets, respectively. The accuracies of the models trained from random
initialization were almost similar to those (95.5% and 97.6%, respectively [15]) of the
models trained from transfer learning for the OCT and chest X-ray image datasets; however,
the accuracy from random initialization was slightly lower than that (87.7% [15]) from
transfer learning.

We evaluated the vulnerability of nontargeted UAPs against these Inception V3 models
(Table 2) and found that the UAPs from natural images were partly less effective for fooling
the DNN-based medical image classifications. For the skin lesion image classification, the
\( R_f \) of the UAP from the ImageNet dataset was only \( \sim 50\% \), despite a larger \( \zeta \) (\( \zeta = 8\% \),
i.e., two times larger than the case show in Fig. 3a), whereas \( R_f \) of the UAP from the
training images was \( \sim 90\% \). For the chest X-ray image classification, \( R_f \) of the UAP from
the ImageNet dataset was only \( \sim 20\% \) despite a larger \( \zeta \) (\( \zeta = 8\% \), i.e., two times larger
than the case shown in Fig. 3c), whereas \( R_f \) of the UAP from the training images was
\( \sim 45\% \). The results indicate that model training from random initialization reduces the
performance of UAPs from natural images. However, the vulnerability of UAPs from
natural images is not completely avoided because of random initialization. \( R_f \) of the
UAPs from the ImageNet dataset was still larger than that of random UPAs, \( R_f \) was
almost similar between the UAPs from the ImageNet dataset and the UAPs from the
training images for the OCT image classification, although \( \zeta = 16\% \), i.e., two times larger
than the case shown in Fig. 3b).

**Discussion**

We hypothesized that UAPs against DNN models with transfer learning are also
generatable using natural images because pre-trained models do not change significantly
after fine-tuning. We further demonstrate that fine-tuned models for medical image
classification are vulnerable to both non-targeted and targeted UAPs generated using
natural images (Fig. 1 and Table 1). Vulnerability to both non-targeted and targeted UAPs
from natural images was confirmed in several model architectures. Vulnerability to UAPs
may be a universal feature in DNNs. Given the fact that the DNN models for medical image
classification with transfer learning from the ImageNet dataset are vulnerable not only to
UAPs from the ImageNet dataset but also to UAPs from the Open Images datasets, this
vulnerability to the UAPs may be independent of natural image datasets, indicating that
UAPs against DNN models with transfer learning are generatable using any publicly-
available natural images. This may be a novel security threat to DNN-based medical image
diagnosis, in particular, it indicates that almost imperceptible UAPs are generatable without
training medical data and any other similar medical data (no matter how much such data
are kept a secret). Unlike the prediction that adversarial attacks on DNN-based medical
image classifications are difficult because data availability of medical images is generally
limited in terms of security and privacy preservation, the results indicate that a DNN-based
medical image diagnosis is easier to deceive. Adversaries can not only result in failed
DNN-based medical image diagnoses but can also control DNN-based medical image
diagnoses, even if they never access such medical data.
The UAPs from natural images seem to be different from those from training images (Figs. 4 and 6) and the characteristics (e.g., $R_f$, dominant classes, and $R_s$) of the UAPs from natural images were partly different from those of the UAPs from training images. This may be because of the difference in the composition of the predicted labels between the training images and natural images (Additional file 1: Tables S1 – S3). For chest X-ray image classification, for example, ~80% of both the ImageNet and Open Images datasets were classified as PNEUMONIA regardless of model architecture (Additional file 1: Table S3), whereas the training images were almost class-balanced. As the nontargeted attack algorithm [16] considers maximizing $R_f$, a large $R_f$ is achieved when images with such an abundant label are misclassified. On the other hand, misclassifying images with a less abundant label has little advantage for maximizing $R_f$. The performance of nontargeted UAPs is less effective (images with less abundant labels are difficult-to-fool) and less-abundant labels tend to correspond to dominant classes when the predicted labels of natural images are imbalanced. For the chest X-ray image classification, the dominant class of the UAPs from natural images was NORMAL (Fig. 5); as a result, $R_f$ was saturated at ~50% (Fig. 1c). The tendency of the dominant classes to correspond to the less-abundant predicted labels (see Tables S1 and S2) was also observed for the skin lesion and OCT image classification (Fig. 5). The imbalanced predicted labels of the natural images also affect the performance of the targeted UAPs. As the targeted attack algorithm [21] considers maximizing $R_s$, a large $R_s$ will have already been achieved for targeted attacks to an abundant label in a dataset. Thus, UAPs are rarely updated in the iterative algorithm; as a result, $R_s$ rarely increases. The targeted attacks to NM and PNEUMONIA, which are the abundant labels in the dataset (Additional file 1: Tables S2 and S3), were less effective for the OCT and chest X-ray image classifications, respectively (Table 2). The performance of UAPs from natural images may increase by controlling the composition of the predicted labels of the natural images (e.g., using data augmentation).

This study showed that the UAPs were generatable without training data. In this context, UAPs from natural images are regarded as black-box attacks. However, UAPs are not complete black-box attacks because they assume a white-box condition: model parameters (e.g., the gradient of the loss function) are accessible. This was because the well-used UAP algorithms [16, 17], which we also used, are limited to the white-box condition. However, this limitation poses few problems for adversaries. As represented by COVID-Net [22], a deep convolutional neural network design intended to detect COVID-19 cases from chest X-ray images, DNN models are often developed as open-source projects by expecting that both researchers and citizens data scientists will accelerate the development of highly accurate yet practical deep learning solutions. Moreover, collaboration among multiple institutions is required to develop DNN models with high diagnostic performance [23] and the distribution of deep learning models has been proposed as an effective alternative to sharing patient data. Even if model architectures and weights are publicly unavailable and the loss gradient is not accessible, they may be estimated [24] because medical DNNs are often developed by fine-tuning existing DNNs, such as VGG, ResNet, and Inception, pre-trained using the ImageNet dataset, as this study considered. Furthermore, DNNs are aimed at real-world usage (e.g., automated support for clinical diagnosis). The assumption that adversaries cannot access DNN models may be unrealistic from these reasons.

Nevertheless, our findings may also be useful for developing black-box attack methods...
that estimate adversarial perturbations using only model outputs (e.g., confidence scores). Several methods for black-box attacks have been proposed [25–27]. Although they are limited to input-dependent adversarial attacks, universal adversarial attacks may be possible under the black-box condition because CNNs are sensitive to the directions of the Fourier basis functions [28]. However, these methods assume the use of domain-specific data (e.g., medical images in the case of medical image classification) that are not included in the training data. Our study indicates that this assumption is not required. Adversaries may be able to perform black-box attacks more easily than previously thought, simply by using natural images instead of domain-specific images.

A solution for avoiding the vulnerability of UAPs from natural images is to train DNN models from random initialization (i.e., without pretrained weights). The performance of UAPs from natural images was overall lower in the DNN model trained with random initialization (Table 2), compared to the DNN models with transfer learning. This might be because the model weights were different from the pretrained weights from the natural images. However, training from random initialization does not completely prevent the vulnerability of UAPs from natural images. As shown in Table 2, the performance of the UAPs was still higher than that of random controls; moreover, it was almost similar to that of UAPs from training images in some cases (e.g., the OCT image classification). In addition, the trade-offs with prediction performance need to be considered. Because transfer learning contributes to faster convergence [18], prediction performance may decrease when training DNN models from random initialization, compared to transfer learning, when considering the same number of training steps (epochs): this solution may be unrealistic in terms of the practical desire to achieve high prediction performance with a lower computational cost.

Adversarial defenses [29] also need to be considered to reduce vulnerability to UAPs. Recent developments in adversarial defenses [30–33] are remarkable; however, comprehensive comparative evaluations [34, 35] have shown that promising defense methods are less effective than reported. A cat-and-mouse game is defending against adversarial attacks [12]. It may be difficult to completely avoid security concerns caused by adversarial attacks. The development and operation of secure, privacy-preserving, and federated DNNs are needed in medical imaging [11].

**Conclusion**

Our study showed that natural images allow universal adversarial attacks on medical image classification using deep neural networks with transfer learning. It has been expected that adversarial attacks are limited because medical images used for training are generally unavailable; however, existing algorithms can generate UAPs using natural images instead of training images. Transfer learning from natural images is a widely used DNN-based medical image classification because the amount of medical image data is often limited. However, the use of transfer learning causes a security hole, thereby reducing the reliability and safety of computer-based disease diagnosis. Our findings demonstrate a novel vulnerability of DNNs to adversarial attacks and may help to increase the security of DNNs; in particular, they are useful for designing the operation strategy of medical DNNs.
List of abbreviations

AKIEC: actinic keratosis/Bowens disease (intraepithelial carcinoma)

BCC: basal cell carcinoma

BKL: benign keratosis (solar lentigo/seborrheic keratosis/lichen planus-like keratosis)

CNV: neovascular membrane and associated subretinal fluid

DF: dermatofibroma

DME: diabetic macular edema with retinal-thickening-associated intraretinal fluid

DNN: deep neural network

DRUSEN: multiple drusen present in early age-related macular degeneration

MEL: melanoma

NM: normal retina with preserved foveal contour and absence of retinal fluid/edema

NV: melanocytic nevus

OCT: optical coherence tomography

ResNet: residual network

UAP: universal adversarial perturbation

VASC: vascular lesion

VGG: visual geometry group

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

All data generated and analyzed during this study are included in this published article and its supplementary information files. The code and data used in this study are available from our GitHub repository: github.com/kztakemoto/Natural_UAP.
Competing interests

The authors declare that they have no competing interests.

Funding

This research was funded by JSPS KAKENHI (grant number 21H03545).

Authors’ contributions

AM and KT conceived and designed the study. AM and HH prepared the data and model. AM and KT coded and performed the experimental evaluations. AM and KT interpreted the results. KT wrote the manuscript. All authors have approved the final manuscript for publication.

Acknowledgments

We would like to thank Editage (www.editage.jp) for the English language editing.
1. 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.

2. 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. 2017;42 December 2012:60–88. doi:10.1016/j.media.2017.07.005.

3. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2016. p. 2818–26. doi:10.1109/CVPR.2016.308.

4. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2016. p. 770–8. doi:10.1109/CVPR.2016.90.

5. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542:115–8. doi:10.1038/nature21056.

6. Kermany DS, Goldbaum M, Cai W, Valentin CCS, Liang H, Baxter SL, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell. 2018;172:1122-1131.e9. doi:10.1016/j.cell.2018.02.010.

7. 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. 2019;1:e271–97. doi:10.1016/S2589-7500(19)30123-2.

8. Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. Int Conf Learn Represent. 2015. http://arxiv.org/abs/1412.6572.

9. Yuan X, He P, Zhu Q, Li X. Adversarial examples: attacks and defenses for deep learning. IEEE Trans Neural Networks Learn Syst. 2019;30:2805–24. doi:10.1109/TNNLS.2018.2886017.

10. Ortiz-Jimenez G, Modas A, Moosavi-Dezfooli S-M, Frossard P. Optimism in the face of adversity: Understanding and improving deep learning through adversarial robustness. 2020. http://arxiv.org/abs/2010.09624.

11. Kaisiss GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. Nat Mach Intell. 2020;2:305–11. doi:10.1038/s42256-020-0186-1.

12. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. Science. 2019;363:1287–9. doi:10.1126/science.aaw4399.
13. Matyasko A, Chau L-P. Improved network robustness with adversary critic. Proc 32nd Int Conf Neural Inf Process Syst. 2018;:10601–10610. http://arxiv.org/abs/1810.12576.

14. Asgari Taghanaki S, Das A, Hamarneh G. Vulnerability Analysis of Chest X-Ray Image Classification Against Adversarial Attacks. In: Understanding and Interpreting Machine Learning in Medical Image Computing Applications. 2018. p. 87–94. doi:10.1007/978-3-030-02628-8_10.

15. Hirano H, Minagi A, Takemoto K. Universal adversarial attacks on deep neural networks for medical image classification. BMC Med Imaging. 2021;21:9. doi:10.1186/s12880-020-00530-y.

16. 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.

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

18. Raghu M, Zhang C, Kleinberg J, Bengio S. Transfusion: Understanding Transfer Learning for Medical Imaging. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, editors. Advances in Neural Information Processing Systems 32. Curran Associates, Inc.; 2019. p. 3347–57. http://papers.nips.cc/paper/8596-transfusion-understanding-transfer-learning-for-medical-imaging.pdf.

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

20. 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.

21. Hirano H, Takemoto K. Simple iterative method for generating targeted universal adversarial perturbations. In: Proceedings of 25th International Symposium on Artificial Life and Robotics. 2020. p. 426–30. http://arxiv.org/abs/1911.06502.

22. 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. doi:10.1038/s41598-020-76550-z.

23. Chang K, Balachandar N, Lam C, Yi D, Brown J, Beers A, et al. Distributed deep learning networks among institutions for medical imaging. J Am Med Informatics Assoc. 2018;25:945–54.

24. Bortsova G, González-Gonzalo C, Wetstein SC, Dubost F, Katramados I, Hogeweg L, et al. Adversarial attack vulnerability of medical image analysis systems: Unexplored factors. Med Image Anal. 2021;73:102141. doi:10.1016/j.media.2021.102141.
25. 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. doi:10.1016/j.cose.2019.04.014.

26. Guo C, Gardner JR, You Y, Wilson AG, Weinberger KQ. Simple black-box adversarial attacks. Proc 36th Int Conf Mach Learn. 2019;:2484–93. http://arxiv.org/abs/1905.07121.

27. Co KT, Muñoz-González L, de Maupeou S, Lupu EC. Procedural noise adversarial examples for black-box attacks on deep convolutional networks. In: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. New York, NY, USA: ACM; 2019. p. 275–89. doi:10.1145/3319535.3345660.

28. Tsuzuku Y, Sato I. On the structural sensitivity of deep convolutional networks to the directions of Fourier basis functions. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. Computer Vision Foundation / {IEEE}; 2019. p. 51–60. doi:10.1109/CVPR.2019.00014.

29. Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations. 2018. https://openreview.net/forum?id=rJzIBfZAb.

30. Zhang H, Yu Y, Jiao J, Xing E, Ghaoui L El, Jordan M. Theoretically principled trade-off between robustness and accuracy. In: Chaudhuri K, Salakhutdinov R, editors. Proceedings of the 36th International Conference on Machine Learning. Long Beach, California, USA: PMLR; 2019. p. 7472–82. http://proceedings.mlr.press/v97/zhang19p.html.

31. 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.

32. Song C, He K, Wang L, Hopcroft JE. Improving the Generalization of Adversarial Training with Domain Adaptation. 7th Int Conf Learn Represent ICLR 2019. 2019. http://arxiv.org/abs/1810.00740.

33. Hwang U, Park J, Jang H, Yoon S, Cho NI. PuVAE: a variational autoencoder to purify adversarial examples. IEEE Access. 2019;7:126582–93. doi:10.1109/ACCESS.2019.2939352.

34. Croce F, Hein M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. Proc 37th Int Conf Mach Learn. 2020. doi:2003.01690.

35. Carlini N, Wagner D. Adversarial examples are not easily detected. In: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security - AISec ’17. New York, New York, USA: ACM Press; 2017. p. 3–14. doi:10.1145/3128572.3140444.
Table 1: Targeted attack success rates $R_s$ (%) of targeted UAPs against Inception V3, ResNet50, and VGG16 models to each target class. $\zeta = 4\%$ for the skin lesions and chest X-ray image classifications, and $\zeta = 8\%$ for the OCT image classification. The column “UAP” indicates which input image set was used to generate the UAP, except for “Random”, which indicates random UAPs.

| Medical images  | Target class | UAP     | Model architecture |
|-----------------|--------------|---------|--------------------|
| Skin lesion     |              |         | Inception V3       |
|                 |              |         | ResNet50           |
|                 |              |         | VGG16              |
| NV              | Training     | 97.9    | 99.2               | 98.7               |
|                 | ImageNet     | 98.8    | 96.2               | 86.6               |
|                 | Open Images  | 99.1    | 94.5               | 86.9               |
|                 | Random       | 64.1    | 70.2               | 73.3               |
| MEL             | Training     | 97.1    | 97.7               | 97.6               |
|                 | ImageNet     | 97.1    | 96.0               | 10.5               |
|                 | Open Images  | 96.6    | 94.5               | 10.4               |
|                 | Random       | 14.5    | 11.8               | 8.8                |
| OCT             | Training     | 98.2    | 99.4               | 98.6               |
|                 | ImageNet     | 98.2    | 99.7               | 92.3               |
|                 | Open Images  | 99.4    | 99.8               | 94.0               |
|                 | Random       | 27.6    | 29.3               | 26.5               |
| CNV             | Training     | 99.3    | 99.7               | 99.9               |
|                 | ImageNet     | 99.2    | 35.5               | 98.3               |
|                 | Open Images  | 99.3    | 48.3               | 96.2               |
|                 | Random       | 26.5    | 26.1               | 25.4               |
| Chest X-ray     | Training     | 99.3    | 99.3               | 99.6               |
|                 | ImageNet     | 97.6    | 100                | 95.7               |
|                 | Open Images  | 97.0    | 99.8               | 94.3               |
|                 | Random       | 55.7    | 54.4               | 54.8               |
| PNEUMONIA       | Training     | 97.8    | 99.1               | 99.8               |
|                 | ImageNet     | 60.0    | 75.3               | 72.3               |
|                 | Open Images  | 62.8    | 79.8               | 68                 |
|                 | Random       | 45.0    | 46.1               | 44.1               |
Table 2: Fooling rates $R_f$ (%) of nontargeted UAPs against Inception V3 models trained from random initialization. $\zeta = 8\%$ for the skin lesions and chest X-ray image classifications, and $\zeta = 16\%$ for the OCT image classification. The column “UAP” indicates which input image set was used to generate the UAP, except for “Random”, which indicates random UAPs.

| UAP / Medical images | Skin lesion | OCT   | Chest X-ray |
|----------------------|-------------|-------|-------------|
| Training             | 92.7        | 74.5  | 45.9        |
| ImageNet             | 50.0        | 75.3  | 22.2        |
| Random               | 7.3         | 9.9   | 0.4         |
Figure legends

**Figure 1:** Vulnerability to nontargeted UAPs. Line plots of the fooling rate \( R_f \) (%) against Inception V3 model versus perturbation magnitude \( \zeta \) (%) for the skin lesion (a), OCT (b), and chest X-ray (c) image classifications. Legend label denotes the input image set used to generate UAPs, except for “Random”, which indicates random UAPs.

**Figure 2:** Clean images and their adversarial examples generated using nontargeted UAPs from the ImageNet dataset, the against Inception V3 model for the skin lesion (a), OCT (b), and chest X-ray (c) image classifications. \( \zeta = 4\% \) in (a) and (c) \( \zeta = 8\% \) in (b). Labels beside the images are the predicted classes. The clean (original) images are correctly classified into their actual labels.

**Figure 3:** Difference in the fooling rate \( R_f \) (%) of the UAPs according to model architectures for skin lesions (a), OCT (b), and chest X-ray (c) image classifications. \( \zeta = 4\% \) in (a) and (c) \( \zeta = 8\% \) in (b). Dashed lines indicate \( R_f \) (%) of random UAPs (random controls).

**Figure 4:** Visualization of nontargeted UAPs generated using training, ImageNet, and Open Images datasets against Inception V3 models for skin lesion, OCT, and chest X-ray image classifications. UAPs are visually emphasized for clarity; in particular, each UAP is scaled by a maximum of 1 and minimum of 0.

**Figure 5:** Normalized confusion matrices for Inception V3 models attacked using nontargeted UAPs from training, ImageNet, Open Images datasets for skin lesions, OCT, and chest X-ray image classifications.

**Figure 6:** Visualization of targeted UAPs generated using training images, ImageNet, and Open Images datasets against Inception V3 models for skin lesion, OCT, and chest X-ray image classifications. UAPs are visually emphasized for clarity; in particular, each UAP is scaled by a maximum of 1 and minimum of 0.
Additional files

Additional file 1: Supplementary tables and figures. (DOCX)
Figures

Figure 1

Vulnerability to nontargeted UAPs. Line plots of the fooling rate $R_f$ (%) against Inception V3 model versus perturbation magnitude $\zeta$ (%) for the skin lesion (a), OCT (b), and chest X-ray (c) image classifications. Legend label denotes the input image set used to generate UAPs, except for “Random”, which indicates random UAPs.
Figure 2

Clean images and their adversarial examples generated using nontargeted UAPs from the ImageNet dataset, the against Inception V3 model for the skin lesion (a), OCT (b), and chest X-ray (c) image classifications. $\zeta = 4\%$ in (a) and (c) $\zeta = 8\%$ in (b). Labels beside the images are the predicted classes. The clean (original) images are correctly classified into their actual labels.

Figure 3

Difference in the fooling rate $R_f$ (%) of the UAPs according to model architectures for skin lesions (a), OCT (b), and chest X-ray (c) image classifications. $\zeta = 4\%$ in (a) and (c) $\zeta = 8\%$ in (b). Dashed lines indicate $R_f$ (%) of random UAPs (random controls).
Figure 4

Visualization of nontargeted UAPs generated using training, ImageNet, and Open Images datasets against Inception V3 models for skin lesion, OCT, and chest X-ray image classifications. UAPs are visually emphasized for clarity; in particular, each UAP is scaled by a maximum of 1 and minimum of 0.
Figure 5

Normalized confusion matrices for Inception V3 models attacked using nontargeted UAPs from training, ImageNet, Open Images datasets for skin lesions, OCT, and chest X-ray image classifications.
| Skin lesion | ImageNet | Open Images |
|-------------|----------|-------------|
| ![Image](image1.png) | ![Image](image2.png) | ![Image](image3.png) |
| ![Image](image4.png) | ![Image](image5.png) | ![Image](image6.png) |
| ![Image](image7.png) | ![Image](image8.png) | ![Image](image9.png) |

**Figure 6**

Visualization of targeted UAPs generated using training images, ImageNet, and Open Images datasets against Inception V3 models for skin lesion, OCT, and chest X-ray image classifications. UAPs are visually emphasized for clarity; in particular, each UAP is scaled by a maximum of 1 and minimum of 0.

**Supplementary Files**

This is a list of supplementary files associated with this preprint. Click to download.

- [supplementarymaterialAMHHKTNaturalUAP.docx](supplementarymaterialAMHHKTNaturalUAP.docx)