Pattern Recognition 110 (2021) 107332

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Understanding adversarial attacks on deep learning based medical image analysis systems

Xingjun Ma[b,1], Yuhao Niu[a,c,1], Lin Gu[d], Yisen Wang[e], Yitian Zhao[f], James Bailey[b], Feng Lu[a,c,*]

[a] State Key Laboratory of VR Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, China
[b] School of Computing and Information Systems, The University of Melbourne, Parkville, VIC 3010, Australia
[c] Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, Beihang University, Beijing, China
[d] National Institute of Informatics, Tokyo 101-8430, Japan
[e] Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
[f] Cixi Instuitue of Biomedical Engineering, Ningbo Institute of Industrial Technology, Chinese Academy of Sciences, Ningbo, China

ARTICLE INFO

Article history:
Received 18 July 2019
Revised 3 March 2020
Accepted 12 March 2020
Available online 1 May 2020

Keywords:
Adversarial attack
Adversarial example detection
Medical image analysis
Deep learning

ABSTRACT

Deep neural networks (DNNs) have become popular for medical image analysis tasks like cancer diagnosis and lesion detection. However, a recent study demonstrates that medical deep learning systems can be compromised by carefully-engineered adversarial examples/attacks with small imperceptible perturbations. This raises safety concerns about the deployment of these systems in clinical settings. In this paper, we provide a deeper understanding of adversarial examples in the context of medical images. We find that medical DNN models can be more vulnerable to adversarial attacks compared to models for natural images, according to two different viewpoints. Surprisingly, we also find that medical adversarial attacks can be easily detected, i.e., simple detectors can achieve over 98% detection AUC against state-of-the-art attacks, due to fundamental feature differences compared to normal examples. We believe these findings may be a useful basis to approach the design of more explainable and secure medical deep learning systems.

1. Introduction

Deep neural networks (DNNs) are powerful models that have been widely used to achieve near human-level performance on a variety of natural image analysis tasks such as image classification [1], object detection [2], image retrieval [3] and 3D analysis [4]. Driven by their current success on natural images (eg. images captured from natural scenes such as CIFAR-10 and ImageNet), DNNs have become a popular tool for medical image processing tasks, such as cancer diagnosis [5], diabetic retinopathy detection [6] and organ/landmark localization [7]. Despite their superior performance, recent studies have found that state-of-the-art DNNs are vulnerable to carefully crafted adversarial examples (or attacks), *i.e.*, slightly perturbed input instances can fool DNNs into making incorrect predictions with high confidence [8,9]. This has raised safety concerns about the deployment of deep learning models in safety-critical applications such as autonomous driving [10], action analysis [11] and medical diagnosis [12].

While existing works on adversarial machine learning research have mostly focused on natural images, a full understanding of adversarial attacks in the medical image domain is still open. Medical images can have domain-specific characteristics that are quite different from natural images, for example, unique biological textures. A recent work has confirmed that medical deep learning systems can also be compromised by adversarial attacks [12]. As shown in Fig. 1, across three medical image datasets Fundoscopy [6], Chest X-Ray [13] and Dermoscopy [14], diagnosis results can be arbitrarily manipulated by adversarial attacks. Such a vulnerability has also been discussed in 3D volumetric medical image segmentation [15]. Considering the vast sums of money which underpin the healthcare economy, this inevitably creates risks whereby potential attackers may seek to profit from manipulation against the healthcare system. For example, an attacker might manipulate their examination reports to commit insurance fraud or a false claim of medical reimbursement [16]. On the other hand, an attacker might

* Corresponding author at: State Key Laboratory of VR Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, China.
E-mail addresses: xingjun.ma@unimelb.edu.au (X. Ma), niuyuhao@buaa.edu.cn (Y. Niu), ling@nii.ac.jp (L. Gu), eewangyisen@gmail.com (Y. Wang), yitian.zhao@nimte.ac.cn (Y. Zhao), baileyj@unimelb.edu.au (J. Bailey), lufeng@buaa.edu.cn (F. Lu).
[1] Equal Contribution.

https://doi.org/10.1016/j.patcog.2020.107332
0031-3203/

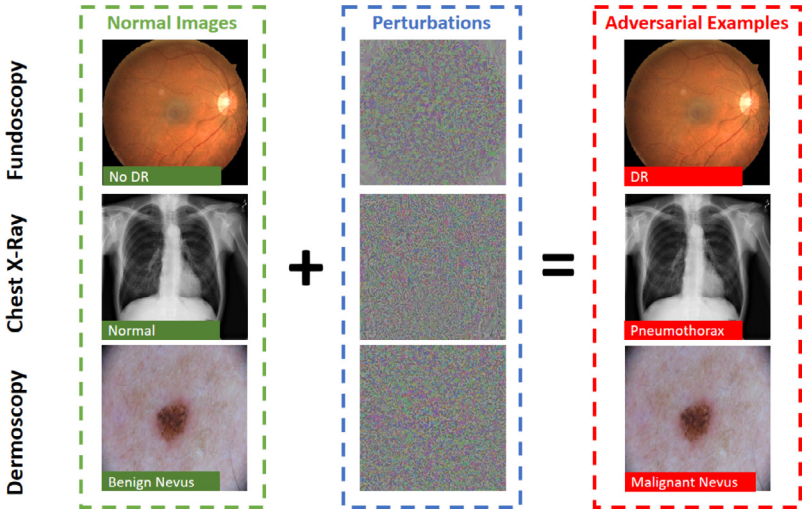

Fig. 1. Examples of adversarial attacks crafted by the Projected Gradient Descent (PGD) to fool DNNs trained on medical image datasets Fundoscopy [6] (first row, DR=diabetic retinopathy), Chest X-Ray [13] (second row) and Dermoscopy [14] (third row). *Left*: normal images, *Middle*: adversarial perturbations, *Right*: adversarial images. The left bottom tag is the predicted class, and green/red indicates correct/wrong predictions.

seek to cause disruption by imperceptibly manipulating an image to cause a misdiagnosis of disease. This could have severe impact for the decisions made about a patient. To make it worse, since the DNN works in a black-box way [17], this falsified decision could hardly be recognised. As deep learning models and medical imaging techniques become increasingly used in the process of medical diagnostics, decision support and pharmaceutical approvals [18], secure and robust medical deep learning systems become crucial [12,16]. A first and important step is to develop a comprehensive understanding of adversarial attacks in this domain.

In this paper, we provide a comprehensive understanding of medical image adversarial attacks from the perspective of generating as well as detecting these attacks. Two recent works [12,16] have investigated adversarial attacks on medical images and mainly focused on testing the robustness of deep models designed for medical image analysis. In particular, the work of [16] tested whether existing medical deep learning models can be attacked by adversarial attacks. They showed that classification accuracy drops from above 87% on normal medical images to almost 0% on adversarial examples. Work in [16] utilized adversarial examples as a measure to evaluate the robustness of medical imaging models in classification or segmentation tasks. Their study was restricted to small perturbations and they observed a marginal but variable performance drop across different models. Despite these studies, the following question has remained open *"Can adversarial attacks on medical images be crafted as easily as attacks on natural images? If not, why?"*. Furthermore, to the best of our knowledge, no previous work has investigated the detection of medical image adversarial examples. A natural question here is to ask *"To what degree are adversarial attacks on medical images detectable?"*. In this paper, we provide some answers to these questions by investigating both the crafting (generation) and detection of adversarial attacks on medical images.

In summary, our main contributions are:

1. We find that adversarial attacks on medical images can succeed more easily than those on natural images. That is, less perturbation is required to craft a successful attack.
2. We show the higher vulnerability of medical image DNNs appears to be due to several reasons: 1) some medical images have complex biological textures, leading to more high gradient regions that are sensitive to small adversarial perturbations; and most importantly, and 2) state-of-the-art DNNs designed for large-scale natural image processing can be overparameterized for medical imaging tasks, resulting in a sharp loss landscape and high vulnerability to adversarial attacks.
3. We show that surprisingly, medical image adversarial attacks can also be easily detected. A simple detector trained on deep features alone can achieve over 98% detection AUC against all tested attacks across our three datasets. To the best of our knowledge, this is the first work on the detection of adversarial attacks in the medical image domain.
4. We show that the high detectability of medical image adversarial examples appears to be because adversarial attacks result in perturbations to widespread regions outside the lesion area. This results in deep feature values for adversarial examples that are recognizably different from those of normal examples.

Our findings of different degrees of adversarial vulnerabilities of DNNs on medical versus natural images can help develop a more comprehensive understanding on the reliability and robustness of deep learning models in different domains. The set of reasons we identified for such a difference reveal more insights into the behavior of DNNs in the presence of different types of adversarial examples. Our analysis of medical adversarial examples provides new interpretations of the learned representations and additional explanations for the decisions made by deep learning models in the context of medical images. This is a useful starting point towards building explainable and robust deep learning systems for medical diagnosis.

The remainder of this paper is organized as follows. In Section 2, we briefly introduce deep learning based medical image analysis. In Section 3, we provide an introduction to adversarial attack and defense techniques. We conduct systematic experiments in Sections 4 & 5 to investigate and understand the behaviour of medical image adversarial attacks. Section 6 discusses several future work and summarizes our contributions.

2. Background of medical image analysis

Driven by the current success of deep learning in traditional computer vision, the field of medical imaging analysis (MIA) has also been influenced by DNN models. One of the first contributions of DNNs was in the area of medical image classification. This

includes several highly successful applications of DNNs in medical diagnosis, such as the severity stage of diabetic retinopathy from retinal fundoscopy [6], lung diseases from chest X-ray [13] or skin cancer from dermoscopic photographs [14]. Another important application of DNNs in medical image analysis is the segmentation of organs or lesions. Organ segmentation aims to quantitatively measure the organs, such as vessels[19,20] and kidneys [21], as a prelude to diagnosis or radiology therapy. Registration is another important task in medical imaging, where the objective is to spatially align medical images from different modalities or capture settings. For example, Cheng et al. [22] exploited the local similarity between CT and MRI images with two types of auto-encoders.

Deep learning based medical image analysis may operate on a variety of input image sources, such as visible light images, hyperspectral light images, X-rays and nuclear magnetic resonance images, across various anatomical areas such as the brain, chest, skin and retina. Brain images have been extensively studied to diagnose Alzheimerâs disease [23] and tumor segmentation [24]. Ophthalmic imaging is another important application, which mainly focuses either on color fundus imaging (CFI) or Optical coherence tomography (OCT) for eye disease diagnosis or abnormalities segmentation. Among these applications, the deep learning based diabetic retinopathy diagnosis system was the first that was approved by the US Food and Drug Administration (FDA). [25] achieved comparable accuracy in detecting diabetic retinopathy to seven certified ophthalmologists using an Inception network. There are systems that apply Convolutional Neural Networks (CNNs) to extract deep features to detect and classify nodules [13] in the chest from radiography and computed tomography (CT). Digital pathology and microscopy is also a popular task due to the heavy burden on clinicians analyzing large numbers of histopathology images of tissue specimens. Specifically, this task involves segmenting high density cells and classifying the mitoses [26]. The above studies rely on the images captured by specialized cameras or devices. In contrast, in the context of skin cancer, it has been shown that standard cameras can deliver excellent performance as input to DNN models [5]. Inspired by this success, the International Skin Imaging Collaboration [14] released a large dataset to support research on melanoma early detection.

Most of these methods, especially diagnosis ones, adopt roughly the same pipeline, on a variety of images including ophthalmology [6], radiology [13] and dermatology [14]. The images are input into CNNs (typically the most advanced ones existing at the time, such as 'AlexNet', 'VGG', 'Inception' and 'ResNet' [1]) to learn intermediate medical features before generating the final output. Whilst these pipelines have achieved excellent success, similar to those for standard computer vision object recognition, they have been criticized for having a lack of transparency. Though some preliminary attempt [17], has been proposed to use Koch postulates, the foundation of evidence based medicine, to explore the decision made by DNNs. People still find it difficult to verify the system's reasoning, which is essential for clinical applications which require high levels of trust. It is easy to see that such trust may be further eroded by the existence of adversarial examples, whereby an imperceptible modification may result in costly and sometimes irreparable damage. We next discuss methods for adversarial attack and detection.

3. Preliminaries

In this paper, we focus on medical image classification tasks using DNNs. For a K-class ($K \geq 2$) classification problem, given a dataset $\{(\mathbf{x}_i, y_i)\}_{i=1,\dots,N}$ with $\mathbf{x}_i \in \mathbb{R}^d$ as a normal example and $y_i \in \{1, \dots, K\}$ as its associated label, a DNN classifier h with parameter $\boldsymbol{\theta}$ predicts the class of an input example $\mathbf{x}_i$:

$$h(\mathbf{x}_i) = \arg\max_{k=1,\dots,K} \mathbf{p}_k(\mathbf{x}_i, \boldsymbol{\theta}), \tag{1}$$

$$\mathbf{p}_k(\mathbf{x}_i, \boldsymbol{\theta}) = \exp(\mathbf{z}_k(\mathbf{x}_i, \boldsymbol{\theta})) / \sum_{k'=1}^{K} \exp(\mathbf{z}_{k'}(\mathbf{x}_i, \boldsymbol{\theta})), \tag{2}$$

where $\mathbf{z}_k(\mathbf{x}_i, \boldsymbol{\theta})$ is the logits output of the network with respect to class k, and $\mathbf{p}_k(\mathbf{x}_i, \boldsymbol{\theta})$ is the probability (softmax on logits) of $\mathbf{x}_i$ belonging to class k. The model parameters $\boldsymbol{\theta}$ are updated using back-propagation to minimize the classification loss such as the commonly used cross entropy loss $\ell(h, \mathbf{x}) = \frac{1}{N}\sum_i^N -y_i \log \mathbf{p}_{y_i}(\mathbf{x}_i, \boldsymbol{\theta})$.

3.1. Adversarial attacks

Given a pretrained DNN model h and a normal sample $\mathbf{x}$ with class label y, an attacking method is to maximize the classification error of the DNN model, whilst keeping $\mathbf{x}_{adv}$ within a small ϵ-ball centered at the original sample $\mathbf{x}$ ($\|\mathbf{x}_{adv} - \mathbf{x}\|_p \leq \epsilon$), where $\|\cdot\|_p$ is the L_p-norm, with L_∞ being the most commonly used norm due to its consistency with respect to human perception [27]. Adversarial attacks can be either targeted or untargeted. A targeted attack is to find an adversarial example $\mathbf{x}_{adv}$ that can be predicted by the DNN to a target class ($h(\mathbf{x}_{adv}) = y_{target}$) which is different from the true class ($y_{target} \neq y$), while an untargeted attack is to find an adversarial example $\mathbf{x}_{adv}$ that can be misclassified to an arbitrary class ($h(\mathbf{x}_{adv}) \neq y$). Adversarial attacks can be generated either in a white-box setting using adversarial gradients extracted directly from the target model, or a black-box setting by attacking a surrogate model or estimation of the adversarial gradients [28,29]. In this paper, we focus on untargeted attacks in the white-box setting under the L_∞ perturbation constraint.

For white-box untargeted attacks, adversarial examples can be generated by solving the following constrained optimization problem:

$$\mathbf{x}_{adv} = \arg\max_{\|\mathbf{x}' - \mathbf{x}\|_\infty \leq \epsilon} \ell(h(\mathbf{x}'), y), \tag{3}$$

where $\ell(\cdot)$ is the classification loss, and y is the ground truth class. A wide range of attacking methods have been proposed for the crafting of adversarial examples. Here, we introduce a selection of the most representative and state-of-the-art attacks.

Fast Gradient Sign Method (FGSM). FGSM perturbs normal examples $\mathbf{x}$ for one step by the amount of ϵ along the input gradient direction [9]:

$$\mathbf{x}_{adv} = \mathbf{x} + \epsilon \cdot \text{sign}(\nabla_{\mathbf{x}} \ell(h(\mathbf{x}), y)). \tag{4}$$

Basic Iterative Method (BIM). BIM [30] is an iterative version of FGSM. Different to FGSM, BIM iteratively perturbs the input with smaller step size,

$$\mathbf{x}^t = \left(\mathbf{x}^{t-1} + \alpha \cdot \text{sign}(\nabla_{\mathbf{x}} \ell(h(\mathbf{x}^{t-1}), y)\right), \tag{5}$$

where α is the step size, and $\mathbf{x}^t$ is the adversarial example at the t-th step ($\mathbf{x}^0 = \mathbf{x}$). The step size is usually set to $\epsilon/T \leq \alpha < \epsilon$ for overall T steps of perturbation.

Projected Gradient Descent (PGD). PGD [27] perturbs a normal example $\mathbf{x}$ for a number of T steps with smaller step size. After each step of perturbation, PGD projects the adversarial example back onto the ϵ-ball of $\mathbf{x}$, if it goes beyond:

$$\mathbf{x}^t = \Pi_\epsilon\left(\mathbf{x}^{t-1} + \alpha \cdot \text{sign}(\nabla_{\mathbf{x}} \ell(h(\mathbf{x}^{t-1}), y))\right), \tag{6}$$

where α is the step size, $\Pi(\cdot)$ is the projection function, and $\mathbf{x}^t$ is the adversarial example at the t-th step ($\mathbf{x}^0 = \mathbf{x}$). Different from BIM, PGD uses random start for $\mathbf{x}^0 = \mathbf{x} + \mathcal{U}^d(-\epsilon, \epsilon)$, where $\mathcal{U}^d(-\epsilon, \epsilon)$ is the uniform distribution between $-\epsilon$ and ϵ, and of the same d dimensions as $\mathbf{x}$. PGD is normally regarded as the strongest first-order attack.

Carlini and Wagner (CW) Attack. The CW attack is a state-of-the-art optimization-based attack [31]. There are two versions of the CW attack: L_2 and L_∞, here we focus on the L_∞ version. According to Madry et al. [27], the L_∞ version of targeted CW attack can be solved by the PGD algorithm iteratively as following

$$\mathbf{x}^t = \Pi_\epsilon\left(\mathbf{x}^{t-1} - \alpha \cdot \text{sign}(\nabla_{\mathbf{x}} \hat{f}(\mathbf{x}^{t-1}))\right) \tag{7}$$

$$\hat{f}(\mathbf{x}^{t-1}) = \max\left(\mathbf{z}_y(\mathbf{x}^{t-1}, \boldsymbol{\theta}) - \mathbf{z}_{y_{max} \neq y}(\mathbf{x}^{t-1}, \boldsymbol{\theta}), -\kappa\right), \tag{8}$$

where $\hat{f}(\cdot)$ is the surrogate loss for the constrained optimization problem defined in Eq. (3), $\mathbf{z}_y$ is the logits with respect to class y, $\mathbf{z}_{y_{max} \neq y}$ is the maximum logits of other classes, and κ is a parameter controls the confidence of the attack.

While there also exists other attacking methods [29], in this paper, we focus on the four state-of-the-art attacks mentioned above: FGSM, BIM, PGD and CW.

3.2. Adversarial Detection

A number of defense models have been developed, input denoising [32], input gradients regularization [33], and adversarial training [9,27]. However, these defenses can generally be evaded by the latest attacks, either wholly or partially [34].

Given the inherent challenges for adversarial defense, recent works have instead focused on detecting adversarial examples. These works attempt to discriminate adversarial examples (positive class) from normal clean examples (negative class), based on features extracted from different layers of a DNN. In machine learning, the subspace distance of the high dimension features has long been analysed [35]. Specifically, for the adversarial examples detection, detection subnetworks based on activations [36], a logistic regression detector based on KD and Bayesian Uncertainty (BU) features [37] and the Local Intrinsic Dimensionality (LID) of adversarial subspaces [38] are a few such works.

Kernel Density (KD): KD assumes that normal samples from the same class lie densely on the data manifold while adversarial samples lie in more sparse regions off the data submanifold. Given a point $\mathbf{x}$ of class k, and a set of training samples from the same class X_k, the Gaussian Kernel Density of $\mathbf{x}$ can be estimated by:

$$\text{KD}(\mathbf{x}) = \frac{1}{|X_k|} \sum_{\mathbf{x}' \in X_k} \exp\Big(\frac{\|\mathbf{z}(\mathbf{x}, \boldsymbol{\theta}) - \mathbf{z}(\mathbf{x}', \boldsymbol{\theta})\|_2^2}{\sigma^2}\Big), \tag{9}$$

where σ is the bandwidth parameter controlling the smoothness of the Gaussian estimation, $\mathbf{z}$ is the logits of input $\mathbf{x}$, and $|X_k|$ is the number of samples in X_k.

Local Intrinsic Dimensionality (LID): LID is a measurement to characterize the dimensional characteristics of adversarial subspaces in the vicinity of adversarial examples. Given an input sample $\mathbf{x}$, the MLE estimator of LID makes use of its distances to the first n nearest neighbors:

$$\widehat{\text{LID}}(\mathbf{x}) = -\left(\frac{1}{n}\sum_{i=1}^{n} \log \frac{r_i(\mathbf{x})}{r_n(\mathbf{x})}\right)^{-1}, \tag{10}$$

where $r_i(\mathbf{x})$ is the Euclidean distance between $\mathbf{x}$ and its i-th nearest neighbor, i.e, $r_1(\mathbf{x})$ is the minimum distance while $r_n(\mathbf{x})$ is the maximum distance. LID is computed on each layer of the network producing a vector of LID scores for each sample.

3.3. Classification tasks, datasets and DNN models

Here, we consider three highly successful applications of DNNs for medical image classification: 1) classifying diabetic retinopathy (a type of eye disease) from retinal fundoscopy [39]; 2) classifying thorax diseases from Chest X-rays [13]; and 3) classifying melanoma (a type of skin cancer) from dermoscopic photographs [5]. Here, we briefly introduce some general experimental settings with respect to the datasets and network architectures.

Table 1
Number of classes and images in each subset of the five datasets.

Dataset	Classes	*Train*	*Test*	
			AdvTrain	*AdvTest*
Fundoscopy	2	75397	8515	2129
Chest X-Ray	2	53219	6706	1677
Dermoscopy	2	18438	426	107
Chest X-Ray-3	3	54769	9980	
Chest X-Ray-4	4	57059	10396	

Datasets. We use publicly available benchmark datasets for all three classification tasks. For our model training and attacking experiments, we need two subsets of data for each dataset: 1) subset *Train* for pre-training the DNN model, and 2) subset *Test* for evaluating the DNN models and crafting adversarial attacks. In the detection experiments, we further split the *Test* data into two parts: 1) *AdvTrain* for training adversarial detectors, and 2) *AdvTest* for evaluating the adversarial detectors. The number of classes and images we retrieved from the public datasets can be found in Table 1.

We follow the data collection process described in [12]. For the diabetic retinopathy (DR) classification task, we use the Kaggle dataset Fundoscopy [6], which consists of over 80,000 high-resolution retina images taken under a variety of imaging conditions where each image was labeled to five scales from 'No DR' to 'mid/moderate/severe/proliferative DR'. In accordance with [12,39], we aim to detect the *referable* (grade moderate or worse) diabetic retinopathy from the rest (two classes in total).

For the thorax disease classification task, we use a Chest X-Ray database [13], which comprises 112,120 frontal-view X-ray images of 14 common disease labels. Each image in this dataset can have multiple labels, so we randomly sample images from those labeled only with 'no finding' or 'pneumothorax' to obtain our 2-class dataset. We also sample two multi-class datasets from Chest X-Ray: 1) a 3-class dataset (eg. Chest X-Ray-3 in Table 1) including image labeled only with 'no finding', 'pneumothorax' or 'mass'; 2) a 4-class dataset (eg. Chest X-Ray-4 in Table 1) including 'no finding', 'pneumothorax', 'mass' and 'nodule'.

For the melanoma classification task, we retrieve melanoma related images of class 'benign' and class 'malignant' (two classes in total) from the International Skin Imaging Collaboration database [14]. Fig. 2 shows two examples for each class of our three 2-class datasets.

DNN Models. For all the five datasets, we use the ImageNet pre-trained ResNet-50 [1] as the base network whose top layer is replaced by a new dense layer of 128 neurons, followed by a dropout layer of rate 0.2, and a K neuron dense layer for classification. The networks are trained for 300 epochs using a stochastic gradient descent (SGD) optimizer with initial learning rate 10^{-4}, momentum 0.9. All images are center-cropped to the size $224 \times 224 \times 3$ and normalized to the range of $[-1, 1]$. Simple data augmentations including random rotations, width/height shift and horizontal flip are used. When the training is completed, the networks are fixed in subsequent adversarial experiments.

4. Understanding adversarial attacks on medical image DNNs

In this section, we investigate 4 different attacks against DNNs trained on five medical image datasets. We first describe the attack settings, then present the attack results with accompanying discussions and analyses.

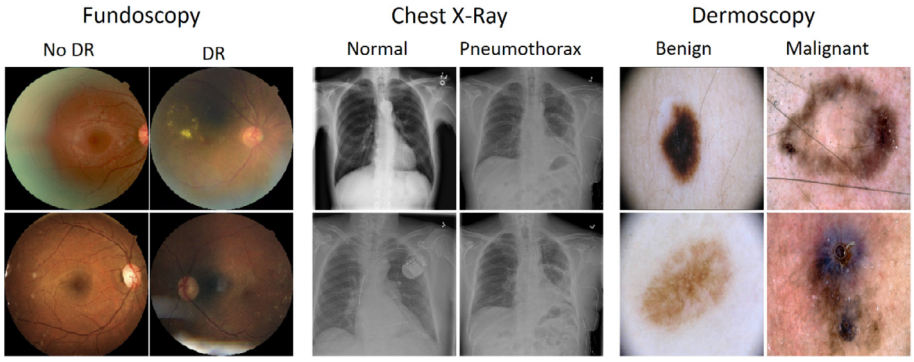

Fig. 2. Example images from each class of the three 2-class datasets.

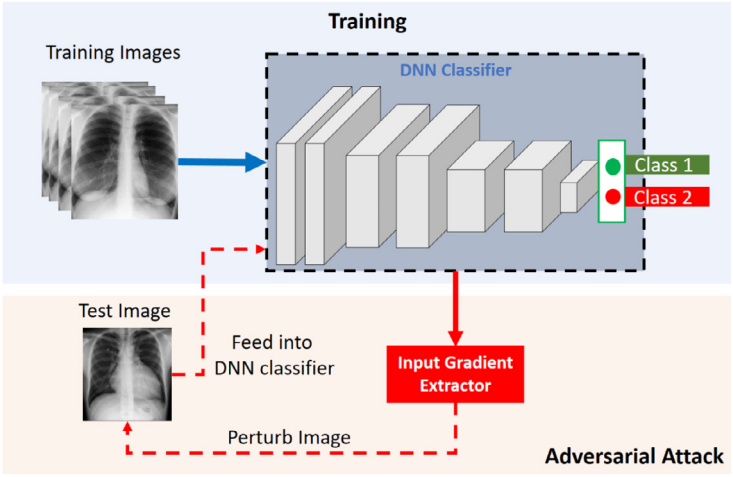

Fig. 3. The pipeline of training DNNs (top) and generating adversarial attacks (bottom).

4.1. Attack settings

The attacks we consider are: 1) the single step attack FGSM, 2) the iterative attack BIM, 3) the strongest first-order attack PGD, and 4) the strongest optimization-based attack CW (L_∞ version). Note that all these attacks are bounded attacks according to a predefined maximum perturbation ϵ with respect to the L_∞ norm, *i.e.*, the maximum perturbation on each input pixel is no greater than ϵ. All 4 types of attacks are applied on both the *AdvTrain* and *AdvTest* subsets of images, following the pipeline in Fig. 3. Given an image, the input gradient extractor feeds the image into the pre-trained DNN classifier to obtain the input gradients, based upon which the image is perturbed to maximize the network's loss to the correct class. The perturbation steps for BIM, PGD and CW are set to 40, 20 and 20 respectively, while the step size are set to $\epsilon/40$, $\epsilon/10$ and $\epsilon/10$ accordingly. We focus on untargeted attacks in a white-box setting.

4.2. Attack results

We focus on the difficulty of adversarial attack on medical images compared to that on natural images in ImageNet. The attack difficulty is measured by the least maximum perturbation required for most (*e.g.* $>$ 99%) attacks to succeed. Specifically, we vary the maximum perturbation size ϵ from 0.2/255 to 5/255, and visualize the drop in model accuracy on the adversarial examples in Figs. 4 and 5 for our 2-class and multi-class datasets respectively, and the numeric results with respect to maximum perturbation $\epsilon = 1.0/255$ can be found in Tables 2 and 3 separately.

Table 2
The classification accuracies (%) and AUCs (%) of the three 2-class DNN classifiers on clean test images (denoted as "No attack") and the 4 types of adversarial examples under L_∞ maximum perturbation 1.0/255.

Attack	Fundoscopy		Chest X-Ray		Dermoscopy	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
No attack	91.03	81.91	93.99	61.25	87.62	78.74
FGSM	1.15	3.71	1.90	0.96	29.98	20.58
BIM	0.00	0.00	0.00	0.00	0.21	0.13
PGD	0.00	0.00	0.00	0.00	0.43	0.74
CW	0.04	0.09	0.00	0.00	0.21	0.13

Table 3
White-box attacks on 2-class versus multi-class models on datasets Chest X-Ray (CXR-2), Chest X-Ray-3 (CXR-3) and Chest X-Ray-4 (CXR-4): the classification accuracies (%) of the three DNN classifiers on clean test images (denoted as "No attack") and the 4 types of adversarial examples under L_∞ maximum perturbation $\epsilon = 0.3/255$ and $\epsilon = 1.0/255$.

Attack	Accuracy when $\epsilon = 0.3/255$			Accuracy when $\epsilon = 1.0/255$		
	CXR-2	CXR-3	CXR-4	CXR-2	CXR-3	CXR-4
No attack	93.99	90.01	84.26	93.99	90.01	84.26
FGSM	16.26	10.07	3.01	1.90	2.14	0.74
BIM	1.60	0.72	0.19	0.00	0.00	0.00
PGD	0.56	0.30	0.08	0.00	0.00	0.00
CW	0.49	1.84	0.17	0.00	0.00	0.00

Results on 2-class datasets. As expected, model accuracy drops drastically when adversarial perturbation increases, similar to that on natural images [9,31]. Strong attacks including BIM, PGD and CW, only require a small maximum perturbation $\epsilon < 1.0/255$ to

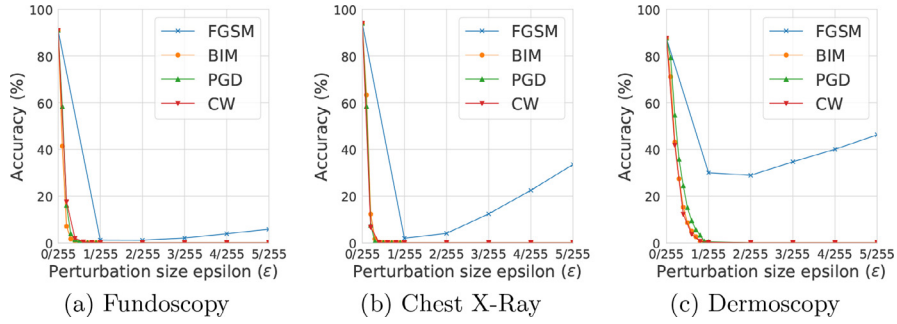

Fig. 4. The classification accuracy of the three 2-class DNN classifiers on adversarial examples crafted by FGSM, BIM, PGD and CW with increasing perturbation size ϵ. Strong attacks including BIM, PGD and CW can succeed most of the time (model accuracy below 1%) with very small perturbation $< 1.0/255$. All attacks were generated in a white-box setting.

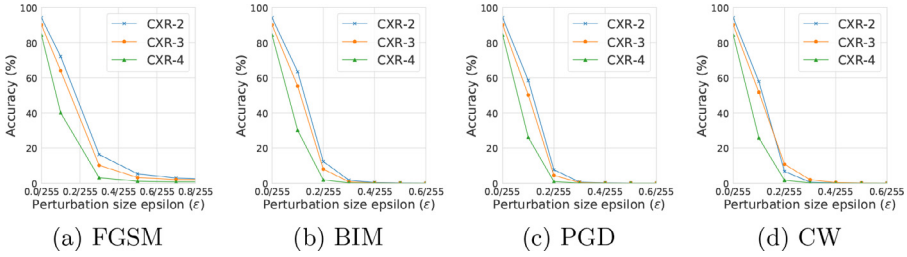

Fig. 5. Comparison of the attacks FGSM, BIM, PGD and CW on datasets Chest X-Ray (CXR-2), Chest X-Ray-3 (CXR-3) and Chest X-Ray-4 (CXR-4). For each attack, the classification accuracy after the attack (in a white-box setting) under different perturbation sizes ϵ is reported.

generally succeed. This means attacking medical images is much easier than attacking natural images like those from CIFAR-10 and ImageNet, which often require a maximum perturbation of $> 8.0/255$ for targeted attacks to generally succeed (see Fig. 2 in [30]).

Results on multi-class datasets. Here, we further investigate the attack difficulty on 2-class datasets (eg. Chest X-Ray) versus that on multi-class datasets (eg. Chest X-Ray-3 and Chest X-Ray-4). As the AUC score is defined with respect to only 2 classes, here we only report the model accuracy on clean images (eg. "No attack") and adversarial images crafted by FGSM, BIM, PGD, and CW. As shown in Table 3, when there are more classes, the attacks have greater success rate. For example, under the same perturbation $\epsilon = 0.3/255$, model accuracy on crafted adversarial examples decreases as the number of classes increases. This indicates that medical image datasets that have multiple classes are even more vulnerable than those 2-class datasets. Similar to the 2-class results above, the attacks BIM, PGD and CW can succeed more than 99% of the time with small perturbation $\epsilon = 1.0/255$. This is the case even with smaller perturbation $\epsilon = 0.3/255$, except for the CW attack on Chest X-Ray-3, which succeeds $> 98\%$ of the time. These findings are consistent with those found on natural images, that is, defending adversarial attacks on datasets with more classes (eg. CIFAR-100/ImageNet versus MNIST/CIFAR-10) is generally more difficult [40].

We next consider further why attacking medical images is much easier than attacking ImageNet images. At first sight it is surprising, since medical images have the same size as ImageNet images.

4.3. Why are medical image DNN models easy to attack?

In this part, we provide explanations to the above phenomenon from the following 2 perspectives: 1) the characteristics of medical images; and 2) the characteristics of DNN models used for medical imaging.

Medical Image Viewpoint. We show the saliency map for several images from different classes, for both ImageNet and medical images in the middle row of Fig. 6. The saliency (or attention) map of an input image highlights the regions that cause the most change in the model output, based on the gradients of the classification loss with respect to the input [41]. We can observe that some medical images have significantly larger high attention regions. This may indicate that the rich biological textures in medical images sometimes distract the DNN model into paying extra attention to areas that are not necessarily related to the diagnosis. Small perturbations in these high attention regions can lead to significant changes in the model output. In other words, this characteristic of medical images increases their vulnerability to adversarial attacks. However, this argument only provides a partial answer to the question, as there is no doubt that some natural images can also have complex textures.

DNN Model Viewpoint. We next show that the higher vulnerability of medical DNN models is largely caused by the use of overparameterized deep networks for simple medical image analysis tasks. The third row in Fig. 6 illustrates the representations learned at an intermediate layer of ResNet-50, i.e., the averaged 'res3a_relu' layer output over all channels. Surprisingly, we find that the deep representations of medical images are rather simple, compared to the complex shapes learned from natural images. This indicates that, on medical images, the DNN model is learning simple patterns (possibly those are only related to the lesions) out of a large attention area. However, learning simple patterns does not require complex deep networks. This motivates us to investigate whether the high vulnerability is caused by the use of overparameterized networks, by exploring the loss landscape around individual input samples. Following previous works for natural adversarial images [42], we construct two adversarial directions $\mathbf{g}$ and $\mathbf{g}^{\perp}$, where $\mathbf{g}$ and $\mathbf{g}^{\perp}$ are the input gradients extracted from the DNN classifiers and a set of separately trained surrogate models respectively. We then craft adversarial examples following $\mathbf{x}_{adv} = \mathbf{x} + \epsilon_1\mathbf{g} + \epsilon_2\mathbf{g}^{\perp}$. More specifically, we gradually increase ϵ_1 and ϵ_2 from 0 to 8.0/255, and visualize the classification loss for

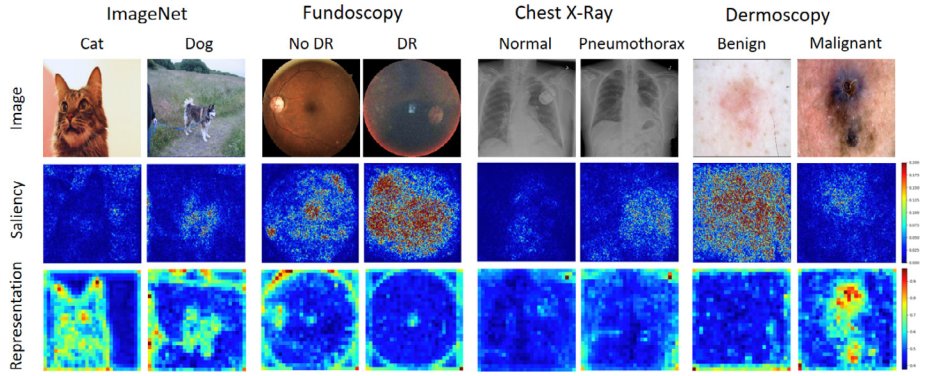

Fig. 6. The normal images (top row), the saliency maps of the images (middle row), and their representations (bottom row) learned at the 'res3a_relu' layer (averaged over channels) of the networks.

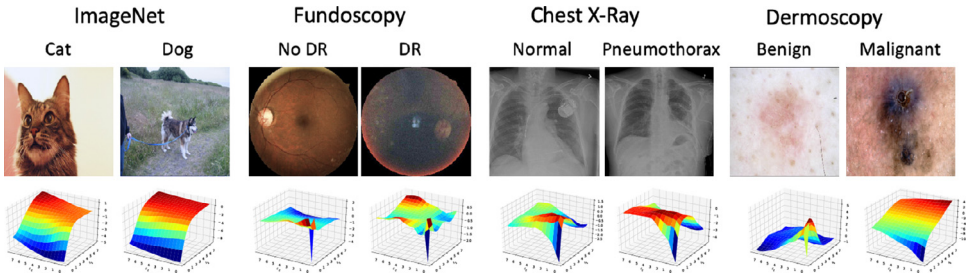

Fig. 7. The landscape (bottom row) of the loss around the input examples (top row). The x, y-axis of the loss landscape plots are ϵ_1 and ϵ_2, which are the sizes of perturbations added to two adversarial directions $\mathbf{g}$ and $\mathbf{g}^{\perp}$ respectively: $\mathbf{x}_{adv} = \mathbf{x} + \epsilon_1\mathbf{g} + \epsilon_2\mathbf{g}^{\perp}$, where $\mathbf{g}$ is the adversarial direction (sign of the input gradients) and $\mathbf{g}^{\perp}$ is the adversarial direction found from the surrogate models. The z-axis of the loss landscape is the classification loss. The use of overparameterized deep networks on medical images causes the loss landscapes around medical images extremely sharp, compared to that of natural images.

each combination of ϵ_1 and ϵ_2 in Fig. 7. We observed that the loss landscapes around medical images are extremely sharp, compared to the flat landscapes around natural images. A direct consequence of sharp loss is high vulnerability to adversarial attacks, because small perturbations of an input sample are likely to cause a drastic increase in loss. A sharp loss is usually caused by the use of an over complex network on a simple classification task [27].

In summary, we have found that medical DNN models can be more vulnerable to adversarial attacks compared to natural image DNN models, and we argue this may be due to 2 reasons: 1) the complex biological textures of medical images may lead to more vulnerable regions; and most importantly, and 2) state-of-the-art deep networks designed for large-scale natural image processing can be overparameterized for medical imaging tasks and result in high vulnerability to adversarial attacks.

4.4. Discussion

In deep learning based medical image analysis, it is a common practice to use state-of-the-art DNNs that were originally designed for complex large-scale natural image processing. However, these networks may be overparameterized for many of the medical imaging tasks. We would like to highlight to researchers in the field that, while these networks bring better prediction performance, they are more vulnerable to adversarial attacks. In conjunction with these DNNs, regularizations or training strategies that can smooth out the loss around input samples may be necessary for robust defenses against such attacks.

5. Understanding the detection of medical image attacks

In this section, we conduct various adversarial detection experiments using two state-of-the-art detection methods, *i.e.*, KD [37] and LID [38]. In addition, we also investigate the use of deep features (denoted by "DFeat") or quantized deep features (denoted by "QFeat") [43] for adversarial detection. The detection experiments are conducted on our three 2-class datasets.

5.1. Detection settings

The DNN models used here are the same as those used in the above attack experiments (see Section 4). The detection pipeline is illustrated in Fig. 8. Based on the pretrained DNN models, we apply the four attacking methods (*FGSM, BIM, PGD and CW*) to generate adversarial examples for the correctly classified images from both the *AdvTrain* and *AdvTest* subsets. We then extract the features used for detection, which include the deep features at the second-last dense layer of the network ("DFeat"/"QFeat"), the KD (kernel density estimated from the second-last layer deep features) features, and the LID (local intrinsic dimensionality estimated from the output at each layer of the network) features. All the parameters for KD/LID estimation are set as per their original papers. All detection features are extracted in mini-batches of size 100. The detection features are then normalized to [0,1]. The detectors are trained on the detection features of the *AdvTrain* subset, and tested on the *AdvTest* subset. As suggested by Feinman et al. [37], Ma et al. [38], we use a logistic regression classifier as the detector for KD and LID, the random forests classifier as the detector for the deep features, and the SVM classifier for quantized deep features. AUC (Area Under Curve) score is adopted as the metric for detection performance.

5.2. Detection results

We report the detection AUC scores of the 4 types of detectors against the 4 types of attacking methods (white-box) across the three datasets in Table 4. State-of-the-art detectors demonstrate very robust performance against these attacks. Especially the

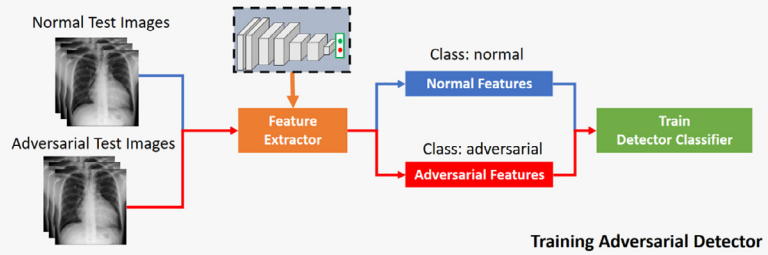

Fig. 8. The pipeline of training an adversarial detector.

Table 4
Detecting white-box attacks: the AUC score (%) of various detectors against the 4 types of attacks crafted on the three datasets. The best results are highlighted in **bold**.

Dataset	Detector	FGSM	BIM	PGD	CW
Fundoscopy	KD	**100.00**	**100.00**	**100.00**	**100.00**
	LID	94.20	99.63	99.52	99.20
	DFeat	99.97	**100.00**	**100.00**	99.99
	QFeat	98.87	99.82	99.91	99.95
Chest X-Ray	KD	99.29	**100.00**	**100.00**	**100.00**
	LID	78.40	96.92	95.20	96.74
	DFeat	**99.97**	**100.00**	**100.00**	**100.00**
	QFeat	87.63	96.35	92.07	99.16
Dermoscopy	KD	**100.00**	**100.00**	**100.00**	**100.00**
	LID	64.83	95.37	92.72	95.90
	DFeat	98.65	99.77	99.48	99.78
	QFeat	86.53	89.27	95.45	93.92

KD-based detectors, which achieve an AUC of above 99% against all attacks across all three datasets. However, on natural images, these state-of-the-art detectors often achieve less than 80% detection AUC against some of the tested attacks such as FGSM and BIM [37,38]. This indicates that medical image adversarial examples are much easier to detect compared to natural image adversarial examples. Quite surprisingly, we find that the deep features (*e.g.* 'DFeat') alone can deliver very robust detection performance against all attacks. In particular, deep feature based detectors achieve an AUC score above 98% across all the testing scenarios. On the other hand, the detectors trained on quantized deep features (*e.g.* 'QFeat') also achieve good detection performance. This indicates that the deep features of adversarial examples (adversarial features) may be fundamentally different from that of normal examples (normal features).

5.3. Detection transferability

We further test if the 'QFeat' detectors can still have good performance when trained on one attack (source), then applied to detect the other 3 attacks (targets). In this transferability test, we train detectors on 'QFeat' of adversarial examples crafted by the source attacks on both *AdvTrain* and *AdvTest* subsets, then apply the trained detectors to detect adversarial examples crafted by other attacks also on both *AdvTrain* and *AdvTest*. As shown in Table 5, the detectors trained on either weak attack FGSM or strong attack PGD all transfer perfectly against other attacks. This again confirms that medical image adversarial examples can be easily detected. The 100% detection AUCs suggests that there are indeed some fundamental differences between adversarial examples and normal examples.

5.4. Why are adversarial attacks on medical images easy to detect?

To better illustrate the difference between adversarial and normal features, we visualize the 2D embeddings of the deep features

Table 5
The detection transferability of the 'DFeat' detector: the AUC score (%) of the two detectors trained on source attacks FGSM and PGD then applied to detect other 3 attacks. The best results are highlighted in **bold**.

Dataset	Source	FGSM	BIM	PGD	CW
Fundoscopy	FGSM	–	**100.00**	**100.00**	**100.00**
	PGD	**100.00**	**100.00**	–	**100.00**
Chest X-Ray	FGSM	–	**100.00**	**100.00**	**100.00**
	PGD	**100.00**	**100.00**	–	**100.00**
Dermoscopy	FGSM	–	**100.00**	**100.00**	**100.00**
	PGD	**100.00**	**100.00**	–	**100.00**

using *t*-SNE [44]. We observe in Fig. 9 that adversarial features are almost linearly separable (after some non-linear transformations) from normal features. This is quite different from natural images, where deep features of adversarial examples are quite similar to that of normal examples, and deep feature based detectors can only provide limited robustness [37,38].

Similar to Fig. 6, we visualize the deep representation of normal and adversarial examples in Fig. 10. Here, we focus on features learned at a deeper layer (eg. the 'res5b_relu' layer of ResNet-50), as we are more interested in the cumulative effect of adversarial perturbations. We find that there are clear differences between adversarial and normal representations, especially for medical images. Compared to natural images, adversarial perturbations tend to cause more significant distortions on medical images in the deep feature space. Considering the difference in deep representations between natural images and medical images (Fig. 6), this will lead to effects that are fundamentally different for natural versus medical images. As the deep representations of natural images activate a large area of the representation map, the adversarial representations that are slightly distorted by adversarial perturbations are not significant enough to be different from the normal representations. However, the deep representations of medical images are very simple and often cover a small region of the representation map. We believe this makes small representation distortions stand out as outliers.

To further understand why tiny changes in deep features can make a fundamental difference, we show the attention maps of both normal and adversarial examples in Fig. 11. We exploit the Gradient-weighted Class Activation Mapping (Grad-CAM) technique [45] to find the critical regions in the input image that mostly activate the the network output. Grad-CAM uses the gradients of a target class, flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the class. As demonstrated in Fig. 11, the attentions of the DNN models are heavily disrupted by adversarial perturbations. On natural images, the attentions are only shifted to less important regions which are still related to the target class. For example, in the 'cat' example, the attention is shifted from the ear to the face of the cat. However, on medical images, the attentions are shifted from the lesion to regions that are completely

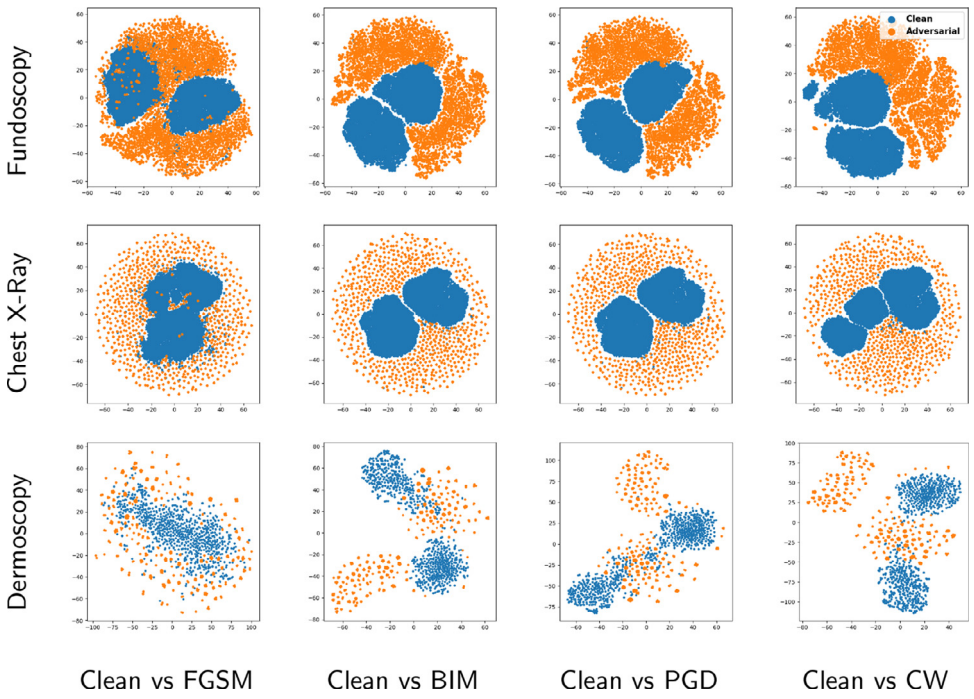

Fig. 9. Visualization of *t*-SNE 2D embeddings of adversarial and normal features, extracted from the second last dense layer of the DNN models. Each row is a dataset, each column is an attack, and blue/orange indicates clean and adversarial examples respectively.

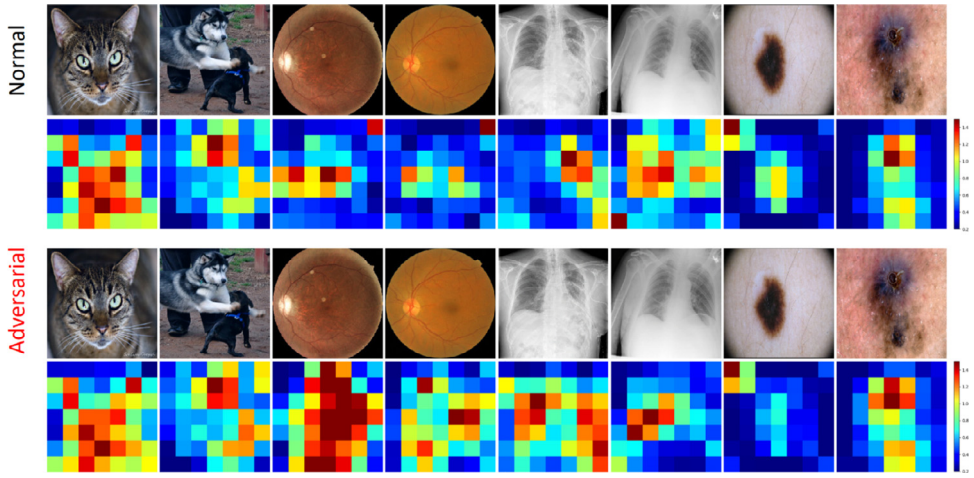

Fig. 10. The deep representations on normal images (first row) versus adversarial images (third row) learned by the ResNet-50 models at the 'res5b_relu' layer (averaged over channels).

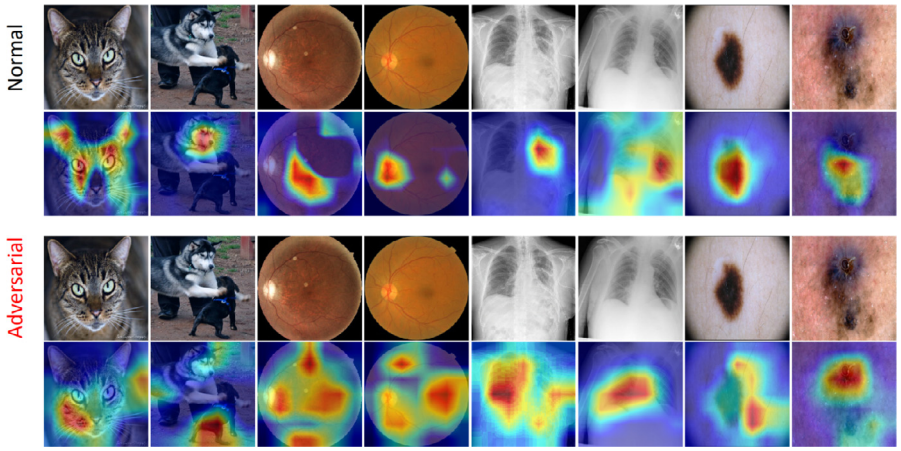

Fig. 11. The attention maps of the network on normal images (first row) versus adversarial images (third row). The attention maps are computed by the Grad-CAM technique [45].

irrelevant to the diagnosis of the lesion. This explains why small perturbations in medical images can lead to deep features that are fundamentally different and easily separable from the normal features.

5.5. Discussion

According to our above analysis, medical image adversarial examples generated using attacking methods developed from natural images are not really "adversarial" from the pathological sense. Careful consideration should be made if using these adversarial examples to evaluate the performance of medical image DNN models. Our study also sheds some light on the future development of more effective attacks on medical images. Pathological image regions might be exploited to craft attacks that produce more misleading adversarial features that are indistinguishable from normal features. Such attacks might have a higher chance to fool both the DNN models and the detectors.

6. Discussion and conclusion

6.1. Discussion

Although existing attacks can easily fool deep neural networks (DNNs) used for medical image analysis, the perturbations are small and imperceptible to human observers, thus posing very limited impact on the diagnosis results when medical experts are involved. Whether physical world medical image examples can be crafted to fool both deep learning medical systems and medical experts is still not clear. While it has been demonstrated possible on natural images [30], traffic signs [10] or object detectors [46], the crafted adversarial stickers or patches are obviously malicious to humans. We believe more subtle and stealthy perturbations will be required for physical-world medical image adversarial examples.

On the defense side, effective defense techniques against medical image adversarial examples are imperative. While existing defense methods developed on natural images such as adversarial training [27,47–49] and regularization methods [33,50] may also apply for medical image adversarial examples, more effective defenses might be developed by also addressing the overparameterization of DNNs used in deep learning medical systems.

6.2. Conclusion

In this paper, we have investigated the problem of adversarial attacks on deep learning based medical image analysis. A series of experiments with 4 types of attack and detection methods were conducted on three benchmark medical image datasets. We found that adversarial attacks on medical images are much easier to craft due to the specific characteristics of medical image data and DNN models. More surprisingly, we found that medical adversarial examples are also much easier to detect, and that simple deep feature based detectors can achieve over 98% detection AUC against all tested attacks across the three datasets and detectors trained on one attack transfer well to detect other unforeseen attacks. This is because adversarial attacks tend to attack a widespread area outside the pathological regions, which results in deep features that are fundamentally different and easily separable from normal features.

Our findings in this paper can help understand why a deep learning medical system makes a wrong decision or diagnosis in the presence of adversarial examples, and more importantly, the difficulties in generating and detecting such attacks on medical images compared to that on natural images. This can further motive more practical and effective defense approaches to improve the adversarial robustness of medical systems. We also believe these findings may be a useful basis to approach the design of more explainable and secure medical deep learning systems.

Acknowledgement

This work was supported by National Natural Science Foundation of China (NSFC) under Grant 61972012 and JST, ACT-X Grant Number JPMJAX190D, Japan and Zhejiang Provincial Natural Science Foundation of China (LZ19F010001).

References

[1] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[2] C. Wang, X. Bai, S. Wang, J. Zhou, P. Ren, Multiscale visual attention networks for object detection in VHR remote sensing images, IEEE Geosci. Remote Sens. Lett. 16 (2) (2019) 310–314.
[3] X. Bai, C. Yan, H. Yang, L. Bai, J. Zhou, E.R. Hancock, Adaptive hash retrieval with kernel based similarity, Pattern Recognit. 75 (2018) 136–148.
[4] F. Lu, X. Chen, I. Sato, Y. Sato, SymPS: BRDF symmetry guided photometric stereo for shape and light source estimation, IEEE Trans. Pattern Anal. Mach.Intell. (TPAMI) 40 (1) (2018) 221–234.
[5] A. Esteva, B. Kuprel, R.A. Novoa, J. Ko, S.M. Swetter, H.M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (7639) (2017) 115.
[6] Kaggle, Kaggle diabetic retinopathy detection challenge, 2015, (https://www.kaggle.com/c/diabetic-retinopathy-detection).
[7] H.R. Roth, L. Lu, A. Farag, H.-C. Shin, J. Liu, E.B. Turkbey, R.M. Summers, DeepOrgan: multi-level deep convolutional networks for automated pancreas segmentation, in: International Conference on Medical Image Computing and Computer Assisted Intervention, Springer, 2015, pp. 556–564.
[8] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, in: International Conference on Learning Representations, 2014.
[9] I.J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, in: International Conference on Learning Representations, 2015.
[10] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, D. Song, Robust physical-world attacks on deep learning visual classification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1625–1634.
[11] Y. Cheng, F. Lu, X. Zhang, Appearance-based gaze estimation via evaluation-guided asymmetric regression, in: European Conference on Computer Vision (ECCV), 2018, pp. 105–121.
[12] S.G. Finlayson, J.D. Bowers, J. Ito, J.L. Zittrain, A.L. Beam, I.S. Kohane, Adversarial attacks on medical machine learning, Science 363 (6433) (2019) 1287–1289.
[13] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R. Summers, ChestX-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3462–3471.
[14] ISIC, The international skin imaging collaboration, 2019, (https://www.isic-archive.com/).
[15] Y. Li, Z. Zhu, Y. Zhou, Y. Xia, W. Shen, E.K. Fishman, A.L. Yuille, Volumetric medical image segmentation: a 3d deep coarse-to-fine framework and its adversarial examples, in: Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics, Springer, 2019, pp. 69–91.
[16] M. Paschali, S. Conjeti, F. Navarro, N. Navab, Generalizability vs. Robustness: investigating medical imaging networks using adversarial examples, in: Medical Image Computing and Computer Assisted Intervention, 2018, pp. 493–501.
[17] Y. Niu, L. Gu, F. Lu, F. Lv, Z. Wang, I. Sato, Z. Zhang, Y. Xiao, X. Dai, T. Cheng, Pathological evidence exploration in deep retinal image diagnosis, in: AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 1093–1101.
[18] H.H. Pien, A.J. Fischman, J.H. Thrall, A.G. Sorensen, Using imaging biomarkers to accelerate drug development and clinical trials, Drug Discov. Today 10 (4) (2005) 259–266.
[19] L. Gu, L. Cheng, Learning to boost filamentary structure segmentation, in: International Conference on Computer Vision, 2015.
[20] B. Liu, L. Gu, F. Lu, Unsupervised ensemble strategy for retinal vessel segmentation, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, 2019, pp. 111–119.
[21] Y. Wang, Y. Zhou, W. Shen, S. Park, E.K. Fishman, A.L. Yuille, Abdominal multi-organ segmentation with organ-attention networks and statistical fusion, Med. Image Anal. (2019).
[22] X. Cheng, L. Zhang, L. Zhang, Deep similarity learning for multimodal medical images, Comput. Methods Biomech. Biomed.Eng. (2015).
[23] S. Liu, S. Liu, W. Cai, S. Pujol, R. Kikinis, D. Feng, Early/diagnosis of Alzheimer's disease with deep learning, in: IEEE International Symposium on Biomedical Imaging (ISBI), 2014, pp. 1015–1018.
[24] B.H. Menze, A. Jakab, S. Bauer, et al., The multimodal brain tumor image segmentation benchmark (brats), IEEE Trans. Med. Imaging 34 (10) (2015) 1993–2024.
[25] V. Gulshan, L. Peng, M. Coram, M.C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, et al., Development and validation of a deep learning algo-

rithm for detection of diabetic retinopathy in retinal fundus photographs, J. Am. Med. Assoc. (2016).
[26] D.C. Ciresan, A. Giusti, L.M. Gambardella, J. Schmidhuber, Mitosis detection in breast cancer histology images with deep neural networks., in: International Conference on Medical Image Computing and Computer Assisted Intervention, vol. 8150, Springer, 2013, pp. 411–418.
[27] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, in: International Conference on Learning Representations, 2018.
[28] L. Jiang, X. Ma, S. Chen, J. Bailey, Y.-G. Jiang, Black-box adversarial attacks on video recognition models, in: ACM International Conference on Multimedia, 2019, pp. 864–872.
[29] D. Wu, Y. Wang, S.-T. Xia, J. Bailey, X. Ma, Skip connections matter: On the transferability of adversarial examples generated with resnets, in: International Conference on Learning Representations, 2020.
[30] A. Kurakin, I.J. Goodfellow, S. Bengio, Adversarial examples in the physical world, in: International Conference on Learning Representations, 2017.
[31] N. Carlini, D. Wagner, Towards evaluating the robustness of neural networks, in: 2017 IEEE Symposium on Security and Privacy, IEEE, 2017, pp. 39–57.
[32] Y. Bai, Y. Feng, Y. Wang, T. Dai, S.-T. Xia, Y. Jiang, Hilbert-based generative defense for adversarial examples, in: IEEE International Conference on Computer Vision, 2019, pp. 4784–4793.
[33] A.S. Ross, F. Doshi-Velez, Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
[34] A. Athalye, N. Carlini, D.A. Wagner, Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples, in: International Conference on Machine Learning, 2018, pp. 274–283.
[35] L. Zhou, X. Bai, X. Liu, J. Zhou, E.R. Hancock, Learning binary code for fast nearest subspace search, Pattern Recognit. 98 (2020) 107040.
[36] J.H. Metzen, T. Genewein, V. Fischer, B. Bischoff, On detecting adversarial perturbations, in: International Conference on Learning Representations, 2017.
[37] R. Feinman, R.R. Curtin, S. Shintre, A.B. Gardner, Detecting adversarial samples from artifacts, in: International Conference on Learning Representations, 2017.
[38] X. Ma, B. Li, Y. Wang, S.M. Erfani, S.N.R. Wijewickrema, G. Schoenebeck, M.E. Houle, D. Song, J. Bailey, Characterizing adversarial subspaces using local intrinsic dimensionality, in: International Conference on Learning Representations, 2018.
[39] V. Gulshan, L. Peng, M. Coram, M.C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, et al., Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs, Jama 316 (22) (2016) 2402–2410.
[40] A. Shafahi, M. Najibi, M.A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L.S. Davis, G. Taylor, T. Goldstein, Adversarial training for free!, in: Advances in Neural Information Processing Systems, 2019, pp. 3353–3364.
[41] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, International Conference on Learning Representations, Workshop Track Proceedings (2014).
[42] F. Tramèr, N. Papernot, I. Goodfellow, D. Boneh, P. McDaniel, The space of transferable adversarial examples, arXiv:1704.03453 (2017).
[43] J. Lu, T. Issaranon, D.A. Forsyth, SafetyNet: detecting and rejecting adversarial examples robustly, in: International Conference on Computer Vision, 2017, pp. 446–454.
[44] L.v.d. Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (Nov) (2008) 2579–2605.
[45] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad–CAM: visual explanations from deep networks via gradient-based localization, in: International Conference on Computer Vision, 2017, pp. 618–626.
[46] A. Liu, X. Liu, J. Fan, Y. Ma, A. Zhang, H. Xie, D. Tao, Perceptual-sensitive GAN for generating adversarial patches, in: AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 1028–1035.
[47] Y. Wang, X. Ma, J. Bailey, J. Yi, B. Zhou, Q. Gu, On the convergence and robustness of adversarial training, in: International Conference on Machine Learning, 2019, pp. 6586–6595.
[48] Y. Wang, D. Zou, J. Yi, J. Bailey, X. Ma, Q. Gu, Improving adversarial robustness requires revisiting misclassified examples, in: International Conference on Learning Representations, 2020.
[49] H. Yu, A. Liu, X. Liu, J. Yang, C. Zhang, Towards noise-robust neural networks via progressive adversarial training, arXiv:1909.04839 (2019).
[50] C. Zhang, A. Liu, X. Liu, Y. Xu, H. Yu, Y. Ma, T. Li, Interpreting and improving adversarial robustness with neuron sensitivity, arXiv:1909.06978 (2019).

Xingjun Ma is a research fellow at the School of Computing and Information Systems, The University of Melbourne. He works in areas of adversarial machine learning, deep learning and its applications.

Yuhao Niu is a PhD student with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University. His research interests include computer vision and medical image analysis.

Lin Gu is a project researcher at National Institute of Informatics (NII), Japan. After PhD graduation from the Australian National University, he worked as a postdoctoral researcher at A*STAR, Singapore. He was also a visiting scholar at Kyoto University, Japan. He works on medical imaging and computational photography.

Yisen Wang is an assistant professor at the Department of Computer Science and Engineering, Shanghai Jiao Tong University. He got his Ph.D. degree from Tsinghua University. His research interests are adversarial machine learning and weakly supervised learning.

Yitian Zhao is an associate professor at the Ningbo Institute of Industrial Technology, Chinese Academy of Sciences, China. He obtained his Ph.D. degree in CS from Aberystwyth University, U.K. His primary research interests lie in 2D/3D image processing, medical image analysis, pattern recognition, computer graphics, and database and knowledge base systems.

James Bailey James Bailey is a professor at the School of Computing and Information Systems, The University of Melbourne.

Feng Lu received the Ph.D. degree from The University of Tokyo. He is currently a Professor with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University. His research interests include computer vision, human-computer interaction and augmented intelligence.