big data and cognitive computing

MDPI

Article

RazorNet: Adversarial Training and Noise Training on a Deep Neural Network Fooled by a Shallow Neural Network

Shayan Taheri, Milad Salem and Jiann-Shiun Yuan *

Department of Electrical and Computer Engineering, University of Central Florida, Orlando, FL 32816-2362, USA
* Correspondence: yuanj@mail.ucf.edu; Tel.: +1-407-823-5719

Received: 19 June 2019; Accepted: 16 July 2019; Published: 23 July 2019

Abstract: In this work, we propose ShallowDeepNet, a novel system architecture that includes a shallow and a deep neural network. The shallow neural network has the duty of data preprocessing and generating adversarial samples. The deep neural network has the duty of understanding data and information as well as detecting adversarial samples. The deep neural network gets its weights from transfer learning, adversarial training, and noise training. The system is examined on the biometric (fingerprint and iris) and the pharmaceutical data (pill image). According to the simulation results, the system is capable of improving the detection accuracy of the biometric data from 1.31% to 80.65% when the adversarial data is used and to 93.4% when the adversarial data as well as the noisy data are given to the network. The system performance on the pill image data is increased from 34.55% to 96.03% and then to 98.2%, respectively. Training on different types of noise can benefit us in detecting samples from unknown and unseen adversarial attacks. Meanwhile, the system training on the adversarial data as well as noisy data occurs only once. In fact, retraining the system may improve the performance further. Furthermore, training the system on new types of attacks and noise can help in enhancing the system performance.

Keywords: adversarial attacks; adversarial perturbations; adversarial training; biometric recognition; convolutional neural networks; data security; deep learning; pill recognition; multiple subnetwork; noise training; transfer learning

1. Introduction

The area of deep/machine learning has shown extreme effectiveness and capability in image classification, object recognition, speech recognition, plagiarism detection, and language translation. The application of this area can range from the information technology to automotive industries. Traditionally, the deep/machine learning (D/ML) algorithmic engines need to be designed under the assumption of training them with similar training and test data distributions. According to this assumption, the test samples will be classified in their correct category. However, this assumption may not always be correct, especially with the presence of intelligent adversaries. This has a severe impact on security-critical applications and products. In this context, attacks are designed in a way to evade machine learning-based detection systems. This means the trained models on fully clean data can be vulnerable to maliciously engineered data. Szegedy et al. generated small perturbations on the images for classification problems and fooled the state-of-the-art deep neural networks [1].

Many of the previously proposed defenses are not effective anymore [2–9], especially due to the emergence of new attacks in this area. In fact, a well-engineered adversarial sample can fool neural networks of different models, different architectures, and trained on different data. This is called cross-model and cross-dataset properties of the defenses. It means the adversarial data can

disrupt classification systems and their algorithms. The engineered data can bring weakness in learned representations and classifiers. They show whether the systems are stable in confronting the perturbations or not. As a result, it is important to determine how a neural network should be trained to make it robust to adversarial samples. These attacks can fool applications of different kinds, such as biometric recognition systems or pharmaceutical/clinical trials. For a biometric recognition system, a face image can be modified in a way to cause gender misclassification, while it looks like its original entity [10]. Similarly, a perturbation into the iris or the fingerprint of an entity can lead to denial of services or unauthorized access. On the other hand, injecting malicious perturbations into the pharmaceutical data can result in performing an act of terrorism or even committing murder.

In this work, we contribute to the area of adversarial example detection with application in biometric and pharmaceutical data. Our contributions can be stated as (1) proposing a system called ShallowDeepNet that includes a shallow and a deep neural network. The shallow neural network is responsible for data preprocessing that is defined as generating adversarial samples in (G-Net). The generated adversarial samples from this network are able to fool a deep neural network. The deep neural network or RazorNet is responsible for understanding data and information as well as detecting adversarial samples (D-Net). Therefore, a serial connection of G-Net and D-Net (G+D Net) helps us to detect many of the unknown and unseen attacks. Leveraging a shallow neural network, an attacker is able to disrupt the deep neural network without having access to its model as well as spending shorter training time. (2) Engaging transfer learning for the application of adversarial examples detection: This is one of the few researches that introduces the concept of transfer learning into the detection of adversarial examples. (3) Using both adversarial training and noise training together in our system: We, for the first time, introduce training on noisy data for making a neural network robust. The noisy images can be helpful in making the neural network robust against unknown and unseen adversarial attacks. Any new adversarial attack tries to inject noise/perturbation into images for the sake of fooling the neural network. Therefore, training the deep neural network on diverse and adequate types of noise possibly makes it robust for future adversarial attacks. (4) Utilizing four different types of noise in improving knowledge of detector networks, namely Additive White Gaussian Noise (AWGN), motion blur, reduced contrast and AWGN, and Perlin noise; (5) generating the adversarial versions of two different types of data, biometric (fingerprint and iris) and pill image using well-known adversarial attacks, namely (a) Fast Gradient Sign Method (FGSM), (b) Jacobian-based Saliency Map Attack (JSMA), (c) DeepFool, (d) Carlini and Wagner (C&W), and (e) Projected Gradient Descent (PGD); and (6) assembling and integrating a comprehensive system consisting of all the discussed elements. This shows the significance of this work in terms of implementation. An example of attacking an image by FGSM is shown in Figure 1.

Figure 1. The figure shows injecting a perturbation from the Fast Gradient Sign Method (FGSM) attack into a sample image from the ten-class Canadian Institute for Advanced Research (CIFAR) dataset. A high amount of perturbation is chosen during the simulation for better visual presentation.

Next, we propose a systematic defense based on leveraging the learned knowledge from a clean unrelated dataset, an adversarial unrelated dataset, a noisy dataset, a related manipulated dataset during training, and all the learned knowledge from the last steps in the detection of adversarial

perturbations. The rest of this paper is organized as follows: We discuss the related works in Section 2. Section 3 presents the background information including all the employed concepts and techniques. The proposed systematic defense against the well-known adversarial attacks is illustrated in Section 4. How the process of decision making on the biometric and pill image data can be improved using transfer learning as well as the learned knowledge from the adversarial and the noisy data is explained. The experimental approach along with the results are provided in Section 5. In Section 6, we discuss how this system can be improved in the future and what its possible limitations are. The conclusion is given in Section 7.

2. Related Work

In this section, the related works are described. The area of fooling neural networks is not necessarily limited to images since it can include other types of data such as words. Accordingly, a method has been proposed by Reference [11] that fools a reading comprehension system by adding sentences to the ends of paragraphs by using crowdsourcing. Another work is random character swaps [12] that breaks the output of neural machine translation systems. A similar method has been proposed [13] that can generate a large number of input sentences through the replacement of a word with its synonym. The authors in References [14–17] showed that having adversarial training can help in holding great promise for learning robust models. The authors in Reference [18] presented an application of a multi-threading mechanism for minimization of the training time through rejection of the unnecessary selection of weights. In Reference [19], the authors proposed SeqGAN, that is a sequence generation framework for solving the problems (a) of difficulty in passing the gradient update from the discriminative model to the generative model and (b) of the limitation of the discriminative model in assessing partially generated sequences. The second problem is similar to the problem of assessing adversarial data. Application of image processing techniques for noise removal can be helpful in overcoming the threats of adversarial examples, for example, taking the architectures for applying multi-frame SR with JPEG2000 compression (working based on a modified adaptive Wiener filter) [20] and leveraging a computer-aided lung nodule detection system into the context of adversarial example detection [21]. Chivukula and Liu show an adversarial learning algorithm for supervised classification, specifically convolutional neural networks [22]. The proposing algorithm has the duty of producing minor changes to the data distribution defined over positive and negative class labels. The work is further augmented by proposing a network capable of defending against unforeseen changes in the data. Kwon et al. [23] proposed a multi-targeted adversarial example that is capable of misclassifying each of the multiple models as each target class along with minimizing the distance of the original sample. A poisoning attack called TensorClog has been proposed in Reference [24] according to which the deep neural networks are jeopardized. The authors in Reference [25] proposed a novel hybrid modular artificial neural network (ANN) architecture that is capable of constructing smooth polygonal meshes from a single depth frame with beforehand knowledge. An investigation on the robustness of the representations learned by the fooled neural network (analyzing the activations of its hidden layers) has been done in Reference [26]. Through this investigation, they tested scoring approaches employed for k-nearest neighbor classifications to distinguish between correctly classified authentic clean images and adversarial images. A defense mechanism for the vulnerability of neural networks to adversarial examples is presented in Reference [27].

In addition to the above publications, a number of the proposed works in this area are based on detection of adversarial examples by relying on adding an outlier class detection module to the classifier [28–30]. A detection model has been presented in Reference [31] that operates based on kernel density estimation and Bayesian neural network uncertainty. A work presented by Reference [32] showed that all the defense methods can be bypassed. Other types of work in this area are based on learning network features and on adapting them to different domains for the same task [33]. Similar works are poisoning attacks that have been mainly explored in the context of binary classification. In a recent work, the vulnerabilities of capsule networks to adversarial attacks (i.e., targeted and untargeted,

black and white box, and individual universal) are studied. It is shown that these attacks, when applied to the German Traffic Sign Recognition Benchmark (GTSRB), are capable of misleading the capsule networks [34].

3. Technical Background

The fundamental concepts and techniques used in this work are discussed. These concepts and techniques include data recognition system for fingerprint, iris, and pill image data; well-known adversarial attacks for fooling neural networks; noise training; and shallow-deep CNN-based system architecture.

3.1. Data Recognition System for Fingerprint, Iris, and Pill Image Data

Here, we discuss two targeted neural network-based systems for adversarial attacks, namely Biometric Recognition System and Pill Recognition System. A scientific medium to distinguish different objects in a reliable manner for a target application based on the physical or behavioral traits of entities (such as fingerprint and iris) is called biometrics. A system for recognition of biometric data tries to find patterns inside the data and to extract features from them to be compared against the reference data. This type of recognition system has many security-related applications, including access control, time, attendance management system, government and law enforcement, passport-free automated border crossings, national ID systems, computer login, and other wireless-based devices for authentication. Two reputable biometric data with significant gained attention are fingerprint and iris. The fundamental and traditional biometric data is fingerprint with having recent applications in smart phones. This trait uses the patterns of ridge and valleys on the surface of a fingertip. Using the print instance from multiple fingers can further enhance the level of security. On the other hand, having small cuts or bruises along with aging and exposure to the environmental disturbances can cause performance degradation of the system. The other trait is iris that has newer applications. It is the annular region of the eye, surrounding the pupil and having sclera on either side. This texture is formed during the fetal development as well and is stabilized as we age. The unique information within this data helps perform recognition and identification tasks. Therefore, having a biometric recognition system with fingerprint and iris image instances as its inputs is one of the best candidates in security provision.

Recently, the usage of prescription drugs has been increased tremendously compared to the past, especially among the elderly. Consequently, the possibility of pill misrecognition has increased significantly. The misrecognition can happen due to the similarities in colors, shapes, imprints, and scorings of the pills. A typical pill recognition system has two modes of learning mode and recognition mode. In the learning mode, we have pill profiling (using the images of pills from a database) and storage. The other mode starts with acquiring the image containing marker and pill. During this process, we have the normalization of image sizes, the detection of markers, the performance of profiling based on the pill shape, size estimation, and color detection. The stored data in the database is used for further consultation. On the other hand, the recognition mode aims for pill detection based on pill profile and feature filtering. The data preprocessing in this mode is similar to the previous mode. Usually, the pill recognition systems are mounted on the mobile devices. In this case, we may have the same issues discussed earlier like low quality images that can lead to misclassification of a pill image. Also, adversarial perturbations can be introduced into the pill images with malicious intent in order to worsen the health condition of a patient. A neural network-based biometric recognition system and a pill recognition system are shown in the top and the bottom parts of Figure 2.

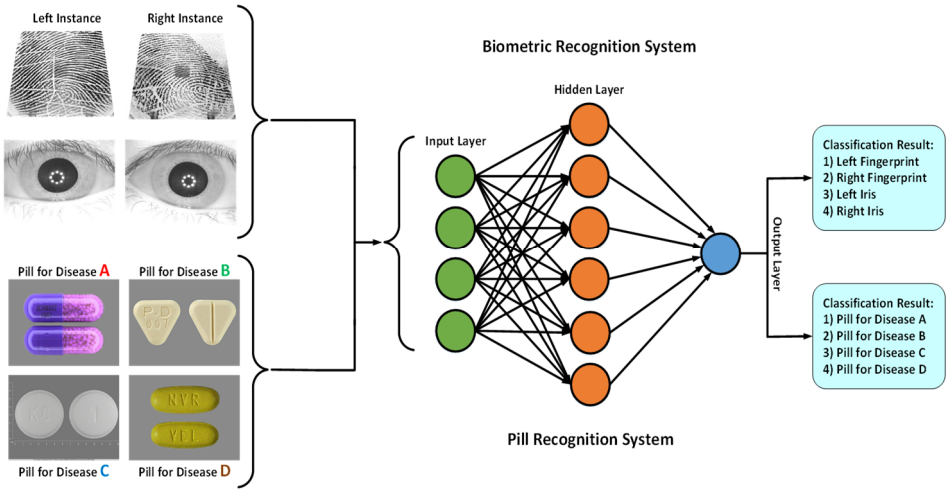


Figure 2. A neural network-based biometric recognition system and a pill recognition system.

3.2. Threat Model: Well-Known Adversarial Attacks for Fooling Neural Networks

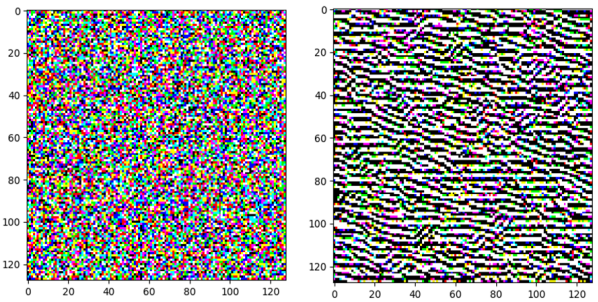
The employed well-known adversarial attacks for fooling neural networks are briefly discussed. The attacks are Fast Gradient Sign Method, Jacobian-Based Saliency Map Attack, DeepFool, Carlini and Wagner Attack, and Projected Gradient Descent. These attacks generate adversarial examples, which are instances of small and intentional feature perturbations that let a machine/deep learning model make a false prediction. These attacks can be modeled by defining F as a classification regime that can output the predicted label $F(x)$ for a given data sample x. Generation of a perturbation R specific to the data sample can cause misclassification based on the equation of $F(x + r) \neq F(x)$. In this perturbation, r should not be distinct enough to be perceived by human beings. Sample instances from this attack are shown in Figure 3.

Fast Gradient Sign Method: FGSM is a fast method for generating adversarial examples [35]. Using this technique, a one-step gradient update is performed along the direction of gradient at each pixel.

Jacobian-Based Saliency Map: A JSMA attack is an efficient saliency adversarial map under L_0 distance [35]. In this attack, a Jacobian matrix is computed with a given sample X and is expressed as $J_f(x) = \frac{\partial f(x)}{\partial x} = \left[\frac{\partial f_i(x)}{\partial x_i}\right]_{i \times j}$.

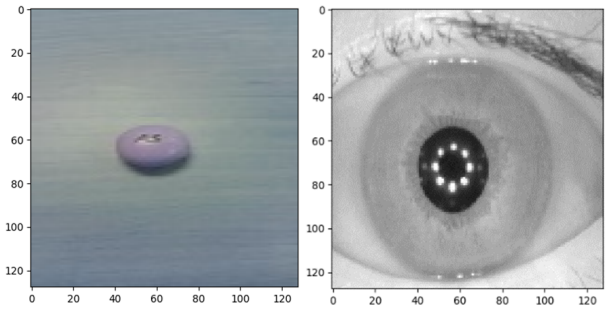
DeepFool: This attack finds the closest distance from the original input to the decision boundary of adversarial samples [35].

Carlini and Wagner: A targeted C&W attack has been offered by Reference [35] for the purpose of defeating the defensive distillation. This attack can bypass most of the existing adversarial detecting defenses.

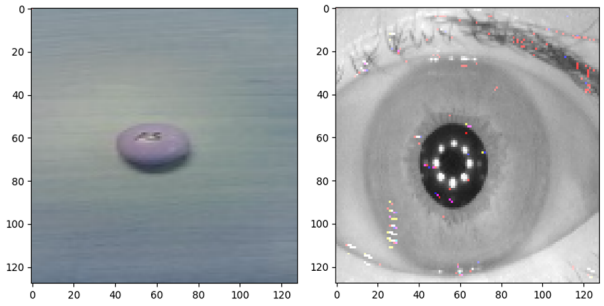


(**a**) FGSM attack

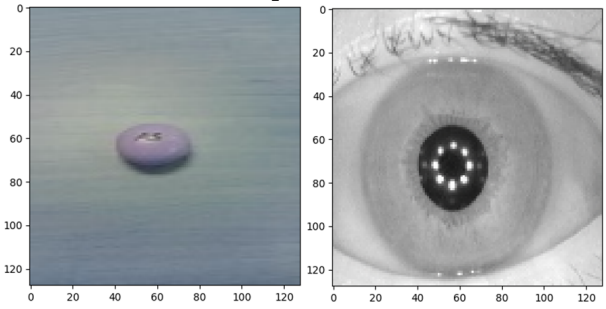
Figure 3. *Cont.*



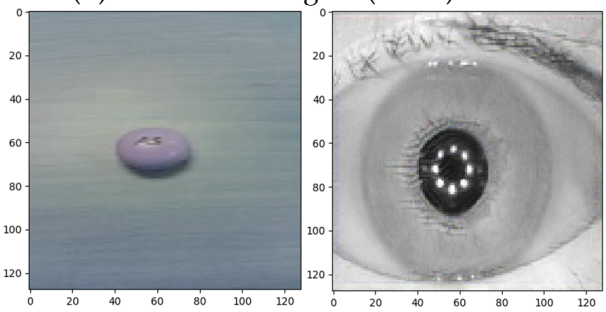
(**b**) Jacobian-based Saliency Map Attack (JSMA)



(**c**) DeepFool attack



(**d**) Carlini and Wagner (C&W) attack



(**e**) Projected Gradient Descent (PGD) attack

Figure 3. Sample instances for (**a**) an FGSM attack; (**b**) JSMA; (**c**) DeepFool attack; (**d**) C&W attack; and (**e**) PGD attack: For each attack, there is pill image data on the left side and the iris biometric data on the right side.

Projected Gradient Descent: A PGD attack model for generating adversarial example has been proposed in Reference [36] according to which the objective problem of $max_{\delta} \leq_{\varepsilon} L(\theta, x + \delta, t_{true})$ is solved.

The parameters used for fooling the shallow neural network are selected based on making the images of original and fooling data look similar to the human eye and have small difference based on the distance measures among the images. In this way, their level of sneakiness will be more difficult to catch using the ordinary defense methods. What we chose for the epsilon perturbation parameter is 0.1 for FGSM, 10 for JSMA, 10 DeepFool, 10 for C&W, and 0.1 for PGD. The attacks are run for 4 epochs.

We performed distance measurements on the original and the perturbed images from the biometric and the pill image datasets. The distance measures are the Euclidean, Manhattan, and Chebyshev distances and the correlation coefficient. Their formulas are presented in Equations (1)–(4). Also, the average of these distances among 100 instances of the original and the adversarial images are shown in Table 1. The method of averaging can be described as (a) finding the distance value between each pair of original and perturbed image; (b) constructing an array of calculated distances; and (c) calculating the average value of the constructed array.

$$\text{Euclidean Distance} : d(x,y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2} \tag{1}$$

$$\text{Manhattan Distance} : d(x,y) = \sum_{i=1}^{m}\left|x_i - y_i\right| \tag{2}$$

$$\text{Chebyshev Distance} : d(x,y) = \max_{i=1,\ldots,m}\left(\left|x_i - y_i\right|\right) \tag{3}$$

$$\text{Correlation Coefficient} : d(x,y) = \frac{\sum_{i=1}^{m}(x_i - \bar{x}).(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{m}(x_i - \bar{x})^2 . \sum_{i=1}^{m}(y_i - \bar{y})^2}} \tag{4}$$

We carry out another statistical analysis in order to show the difference between the fingerprint, iris, and pill image samples. Using this analysis, we can determine the extensibility of our system in recognizing different types of images as well as in defending in front of their adversarial versions. In other words, if there is a unique pattern and similarity among the images used in our experiment, then we cannot determine the system strength because it was successful only in recognizing a certain pattern of data and injections of perturbations inside that specific pattern. In this regard, we run correlation analyses on the iris, fingerprint, and pill image datasets, shown in Figure 4. As it can be seen from the plots in this figure, the correlation coefficients among these images is low enough (below 0.5) to determine that they are not related. In fact, we can say the patterns of iris, fingerprint, and pill are not related. Therefore, it can be said that if our proposing defense system demonstrates perfect performance in recognition and adversarial detection for each of these type of data separately, then the system is extensible and can show a strong performance in recognizing other types of data and detection of adversarial examples for that dataset.

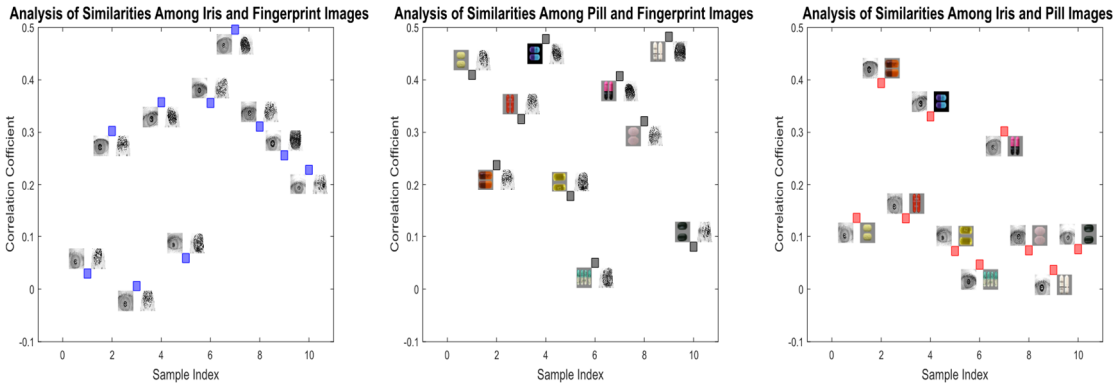


Figure 4. The analysrs of similarities among the images of fingerprint, iris, and pill datasets.

Table 1. The average of distance measures between the original images, perturbed images from biometric, and pill image datasets.

Type of Distance		**Euclidean**			**Manhattan**			**Chebyshev**			**Correlation Coefficient**		
Type of Data		**Fingerprint**	**Iris**	**Pill**	**Fingerprint**	**Iris**	**Pill**	**Fingerprint**	**Iris**	**Pill**	**Fingerprint**	**Iris**	**Pill**
Type of Attack	FGSM	0.12	0.01	0.10	0.37	0.08	0.24	0.11	0.05	0.10	0.99	0.50	0.33
	JSMA	1.86×10^{-4}	8.15×10^{-5}	6.52×10^{-5}	4.88×10^{-4}	2.20×10^{-4}	1.63×10^{-4}	8.55×10^{-5}	3.05×10^{-5}	3.26×10^{-5}	1.0	0.50	0.33
	DeepFool	0.08	0.01	0.05	0.21	0.05	0.11	0.06	0.01	0.04	1.00	0.50	0.33
	C&W	0.02	0.01	0.05	0.11	0.05	0.11	0.01	0.001	0.04	1.00	0.50	0.33
	PGD	0.11	0.01	0.10	0.38	0.10	0.28	0.10	0.01	0.10	0.99	0.49	0.31

3.3. Noise Training

One of the methods in improving the system performance is the addition of noise to the input data of a neural network when it is under training. It has been shown that training a network with noise can be realized as a form of regularization according to which an extra term is added to the error function. The process is performed based upon mixing the noise segments with the original training data [37–39]. The types of noise employed in our work are additive white Gaussian noise (AWGN), motion blur, reduced contrast enhanced AWGN, and Perlin noise. The noisy data helps us learn more information and features that can make the network more knowledgeable to distinguish the difference between clean, adversarial, and noisy data. In fact, its randomness and diversity can be the reason it can detect samples from unknown and unseen attacks.

3.4. Shallow-Deep CNN-Based System Architecture

The authors in Reference [40] proposed a system architecture for the diagnosis of breast cancer according to which the relationships between low energy and recombined images will be discovered. The architecture is capable of applying full field digital mammography for rendering "virtual" recombined images. The classification models have the functionality of performing diagnoses. In simple words, the shallow CNN has the duty of "image reconstruction", and the deep CNN has the duty of "feature extraction". Considering two parallel paths of (a) entering images to a shallow CNN for image reconstruction and giving the output to a deep CNN for feature extraction and (b) using a deep CNN for feature extraction, the features from these paths are combined before determination of the "benign" and the "cancer" image samples. We use a similar idea in our work with the goal strengthening detection of adversarial examples.

We use a similar architecture in the domain of adversarial examples detection—a deep neural network for detection of the adversarial examples generated by the shallow neural network. The concept of Razor is that the supply voltage is tuned for monitoring the error rate during operation [41]. The error detection provides in situ monitoring of the actual circuit delay. This technique relies on a mixture of architectural and circuit level techniques for efficient and effective error detection and correction of delay path failures. This concept can be practically described as augmenting each flip-flop with a so-called shadow latch or Razor latch, controlled by a delayed clock. The Razor latch corrects any error in operation of the main flip-flop since it holds the correct data. According to this concept, we can call the deep neural network RazorNet in our system architecture and it has the duty of detecting error/adversarial samples generated by the shallow neural network.

4. Proposed System and Methodology

We propose a system for the detection of adversarial samples based on three main ideas of (a) shallow-deep system architecture; (b) transfer learning; and (c) adversarial and noisy data training. Two other ideas that can be incorporated into this architecture are retraining on the existing and the future adversarial and noisy data as well as running noise removal techniques on the testing data. The generated adversarial samples along with clean data as well as the noisy data samples will be given to deep CNN for detecting malicious activities. The deep neural network has the duty of understanding those adversarial data.

The architecture of our system is shown in Figure 5. This architecture consists of a generator (f_{gen}) and a detector (f_{det}). f_{gen} has the duty of generating data with adversarial features and gets normal data as its input. f_{det} has the duty of detecting new data with adversarial features and gets normal data, adversarial data, and noisy data as its inputs. Therefore, we have a hybrid training set for the detector. This can be formulated as $D_{HTS} = \left\{(X_{Normal}, Y_{Normal}), (X_{Adv.}, Y_{Adv.}), \left(X_{Noisy}, Y_{Noisy}\right)\right\}$. They have different parameters and layers for feature extraction and training in way f_{det} operates stronger than f_{gen}. The detected adversarial data can be fed back to the deep neural network for retraining that causes better classification accuracy. Based on the concept of Razor latch discussed earlier, the

deep neural network is defined as a Razor neural network (or RazorNet), which has the function of detecting the errors generated by the shallow neural network.

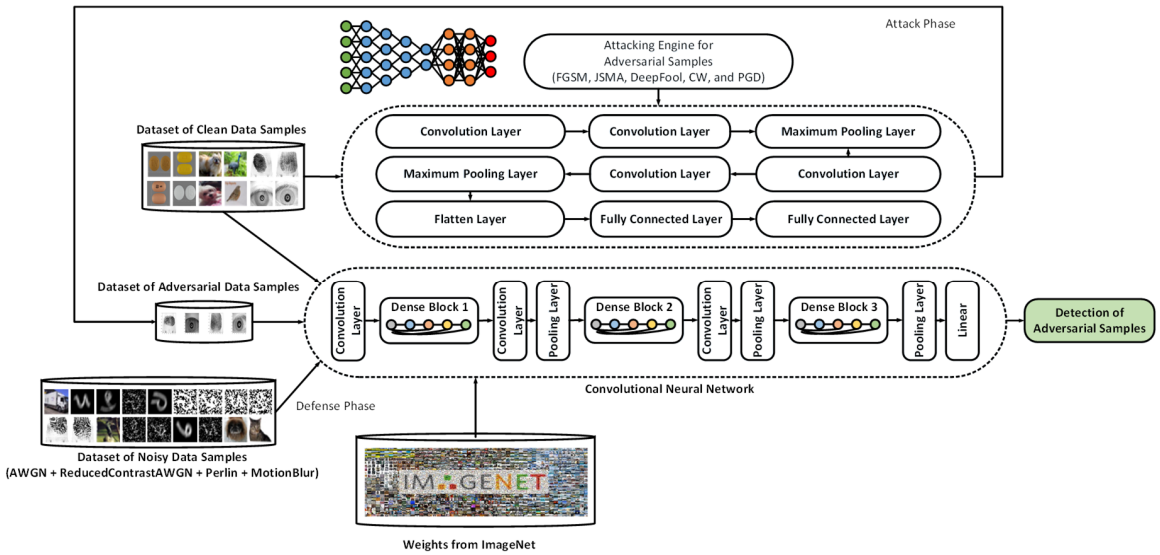


Figure 5. The architecture of our system (G+D Net) for detecting adversarial examples.

The original biometric data as well as the pill images are given to the shallow neural network. The adversarial images generated from this network along with the dataset of clean data, adversarial data, and noisy data samples are input to deep CNN for the detection of malicious patterns. The elements in the architecture of shallow neural network are in the following order: (a) two-dimensional convolution with the filter size of 64 and kernel size of 64; (b) two-dimensional convolution with the filter size of 64, padding of the same, and activation of the rectified linear unit (ReLU); (c) the two-dimensional max pooling with pooling size of (2,2) and stride of 2; (d) two-dimensional convolution with the filter size of 64 and kernel size of 64; (e) two-dimensional convolution with the filter size of 64, padding of the same, and activation of the ReLU; (f) the two-dimensional max pooling with pooling size of (2,2) and stride of 2; (g) flatten layer; (h) two fully connected layers, and (i) a dense layer with four units and one unit for the biometric data (including the left and the right iris and fingerprint data samples) and the pill image data respectively. The deep convolutional neural network is DensetNet with the architecture of a convolution layer, a dense block 1, a convolution layer, a pooling layer, a dense block 2, a convolution layer, a pooling layer, a dense block 3, a pooling layer, and a linear layer. These architectures are summarized in Table 2.

The attacks used in this system are FGSM, JSMA, DeepFool, C&W, and PGD. These attacks try to fool the network with the purpose of causing accuracy drop. The noisy data to be used in this system are (a) additive white Gaussian noise, reduced contrast version of AWGN, and motion blur of the noisy Bangla handwritten digit dataset [42] and (b) added Perlin noise to the CIFAR-10 and Center for Biometrics and Chinese Academy of Sciences' Institute of Automation (CASIA). The generated samples from this network will construct the dataset of adversarial data samples. These samples along with the noisy samples will be used to retrain the deep neural network.

Both the biometric data and the pharmaceutical data are split into the training and the testing sets. This means the system is trained on the training sets and examined on the testing sets. In other words, the system response beyond the training adversarial procedures determines the reported system detection accuracy. Our number of samples for the original biometric data (including iris and fingerprint) is equal to 6664. The ratios of training and testing are 0.8 and 0.2 for this dataset. From the training part of the dataset, a ratio of 0.1 is chosen for validation part of the training. The attacked version of this dataset by the well-known attacks of FGSM, JSMA, DeepFool, C&W, and PGD includes 33,320 image samples.

Table 2. The architectures of the shallow neural network and deep neural network used in our system of adversarial detection.

Shallow Neural Network	Deep Neural Network
Convolution : *Filter* $= 64$; *Kernel* $= (3,3)$; *Same Padding*	*Convolution* : 7×7; *Stride* $= 2$
Convolution : *Filter* $= 64$; *Kernel* $= (3,3)$; *Same Padding*	*Pooling Layer* $= 3 \times 3$; *Stride* $= 2$
Pooling Side $= [2,2]$; *Stride* $= 2$	Dense Block (1) $= \begin{bmatrix} 1\times1 & Conv \\ 3\times3 & Conv \end{bmatrix} \times 6$
Convolution : *Filter* $= 128$; *Kernel* $= (3,3)$; *Same Padding*	1×1 *Convolution*
	Pooling Layer $= 2 \times 2$; *Stride* $= 2$
Convolution : *Filter* $= 128$; *Kernel* $= (3,3)$; *Same Padding*	Dense Block (2) $= \begin{bmatrix} 1\times1 & Conv \\ 3\times3 & Conv \end{bmatrix} \times 12$
Max Pooling; *Pooling Size* $(3,3)$	1×1 *Convolution*
Flatten Layer	*Pooling Layer* $= 2 \times 2$; *Stride* $= 2$
Fully Connected Layer 1	Dense Block (3) $= \begin{bmatrix} 1\times1 & Conv \\ 3\times3 & Conv \end{bmatrix} \times 32$
Fully Connected Layer 2	1×1 *Convolution*
Dense Units = 1/4	*Pooling Layer* $= 2 \times 2$; *Stride* $= 2$
Softmax	Dense Block (4) $= \begin{bmatrix} 1\times1 & Conv \\ 3\times3 & Conv \end{bmatrix} \times 32$
	Pooling Layer $= 7 \times 7$
	Softmax

The pharmaceutical dataset specifically for pill image data is divided into the training and the testing records, each containing 7291 and 800 number of images respectively. The attacked version of these records includes 36,455 and 4000 images, respectively. Our noisy Bangla handwritten digit dataset has 197,889 for each type of noise, including AWGN, motion blur, and reduced contrast and AWGN. Our Perlin dataset comprises 100,000 items (including both CIFAR-10 and CASIA data samples). The retrained network will be able to detect the adversarial examples. The system can be set adaptive in order to adjust its security level for (a) clean data; (b) adversarial data; (c) noisy data; and (d) altogether in different phases in order to adjust itself with respect to the strength of the fooling attacks. In other words, only one case of defense is used if the attack is not strong enough. On the other hand, all cases of the attack can be used if the attack is strong enough to fool the neural network. The Algorithm 1 is shown below.

Algorithm 1: The protocol and overall scheme of the system of shallow-deep neural network architecture, adversarial training, and transfer learning in detection of adversarial perturbations.

01: **Input:** Dataset of clean data samples (X), dataset of noisy data samples (Y), weights from Imagenet (W), shallow neural network model (SM), and deep neural network model (DM)
02: **Output:** Detection of adversarial samples
03: K ← AdversarialSampleGenerator(X, SM)
04: AdversarialSampleDetection ← AdversarialSampleDetector(X, Y, W, K, DM)

5. Experimental Results and Evaluation

In order to evaluate the effectiveness of our architectural model, we used two datasets of CASIA biometric data and 1k Pharmaceutical Pill Image Dataset [43–45]. The real biometric data are chosen from the biometric dataset, and all images of the pill image dataset are used as the clean data. For all these data, their adversarial versions are generated using the attacking engine (which includes FGSM,

JSMA, DeepFool, C&W, and PGD attack functions). Before inputting the adversarial data into the deep neural network, we provide ImageNet weights to the network. This provides an initial knowledge to the network. Besides the adversarial data, the noisy data (AWGN, motion blur, reduced AWGN, and Perlin) are given to the neural network to further augment its understanding and make it capable of distinguishing the clean, adversarial, and noisy data. As it was mentioned earlier, the noisy data can strengthen RazorNet to possibly detect samples from unseen and unknown adversarial attacks. In fact, the network is retrained in this step due to the given weights to the network. The images are all resized with the shape of $32 \times 32 \times 3$ (width $\times$ height $\times$ channel).

The library used in the implementation of our system is the Keras machine learning library. Also, we used Scikit-learn for getting the performance parameters. The deep neural network is pretrained with ImageNet weights and retrained on the adversarial and the noisy data only once [46]. The retraining has been done for five epochs with the shuffled data. The optimizer employed in our experiment is stochastic gradient descent (SGD) [47]. The learning rate is 0.01, the decay is 10^{-6}, and the momentum is 0.9.

According to the simulation results on the biometric data, the system is capable of detecting the adversarial data with 80.65% accuracy when the adversarial data are given into the network and it goes up to 93.4% when both adversarial data and clean data are given into the network. For the pill image data, the system accuracy is improved from 34.55% to 96.03% when the adversarial data is input to the network and from 96.03% to 98.20% when the adversarial data as well as the noisy data are input to the network. In order to make sure that the results are generalizable, we performed five rounds of simulation. According to the simulation results, the system performs completely the same in these runs. Applying the Friedman test on the system outputs will acknowledge this statement. Having this amount of improvement in the results is not out of sight due to the presence of multiple effective components (i.e., adversarial training, noise training, transfer learning, and stronger detection network in terms of the number of layers) in our system. Another reason for the quality of our system is its excellent operation on the datasets from two different domains. In fact, there is no similarity between the biometric and pill image data. Having high performance on these unrelated datasets proves the strength and generalizability of our defense strategy. Meanwhile, the defense system is extensible to other types of data based on the discussion that was provided earlier regarding independency of iris, fingerprint, and pill image data from each other.

The results from examining our system along with other systems for comparison are shown in Table 3. The proposed system can be further improved when we retrain the neural network using the existing and future adversarial and noisy data. for example, retraining the RazorNet using the samples from the obfuscated gradient, one-pixel, and universal perturbation attacks. Other types of noise to be included during training can be brown noise, salt and pepper noise, black noise, and Cauchy noise.

Table 3. The results of our system in detecting adversarial examples (samples).

	Dataset	**System Detection Accuracy on Clean Data**	**System Detection Accuracy on Attacked Data Without Defense**	**System Detection Accuracy on Attacked Data**	
Ours - Biometric Dataset	Chinese Academy of Sciences' Institute of Automation (CASIA) Dataset—Images of iris and fingerprint data	90.58%	1.31%	Clean Data + Adversarial Data	Clean Data + Adversarial Data + Noisy Data
				80.65%	93.4%
Ours - Pillbox Image Dataset	Pillbox Dataset—Images of pill	99.92%	34.55%	96.03%	98.20%
[48]	CIFAR	92%	10%	86%	
[49]	MNIST	N/A	19.39%	75.95%	
[49]	CIFAR	N/A	8.57%	71.38%	
[50]—ResNet	MNIST	88%	0% (Strongest Attack)	83% (Strongest Attack)	
[50]—VGG	MNIST	89%	36% (Strongest Attack)	85% (Strongest Attack)	
[50]—ResNet	CIFAR	85%	7% (Strongest Attack)	71% (Strongest Attack)	
[50]—VGG	CIFAR	82%	37% (Strongest Attack)	80% (Strongest Attack)	

6. Limitations and Future Work

While initial experiments and simulations offer a very promising defense system for adversarial example generation, this architecture may be vulnerable due to a number of reasons: (a) the emergence of new threats for fooling a neural network can break this system via an obfuscated gradient or one-pixel attack. In fact, when a new type of attack is introduced, this system may not be effective due to lack of knowledge for that specific adversarial data. We can tackle this issue by periodically updating the RazorNet. Meanwhile, training RazorNet on the existing and the future noisy data may reinforce the system to detect samples from unknown and unseen adversarial attacks. (b) Considering only one adversarial data generator network (f_{gen}) for the generation of adversarial samples: This is not sufficient in real-world applications, and it is more potent to include diverse types of network models. It means including multiple (N) adversarial data generator networks with different architectures in our system. Figure 6 shows the system of NG+D Net. (c) For higher levels of extensibility of our system, it is beneficial to engage datasets from different domains and to pretrain the detector with diverse types of weights. (d) In order to increase the knowledgeability of our detector, various types of noise and perturbations can be injected into the selected datasets and can perform noise training on the D-Net. Another possibility for extending this work is employing advanced image processing methods for noise removal into our system architecture to overcome the threats of adversarial examples, especially the ones belonging to unknown and unseen attacks. At last, some of the emerging recognition system architectures can be examined in the domain of securing neural networks, namely bilinear CNN [51], gated Siamese CNN architecture [52], HyperFace [53], and EndoNet [54].

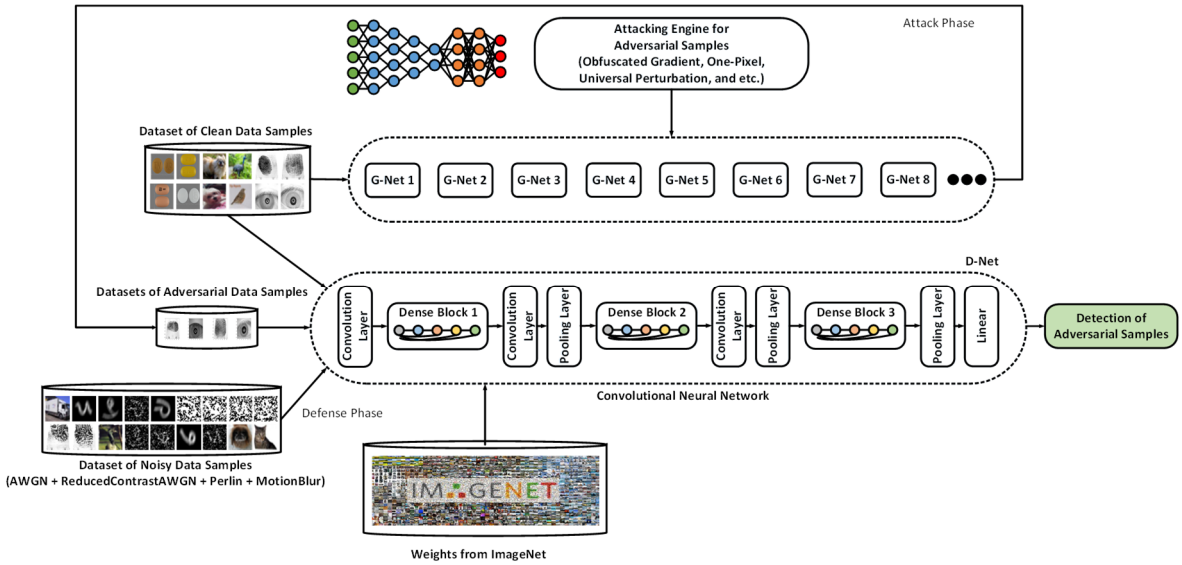


Figure 6. The diagram of the NG+D Net system for detecting adversarial examples.

7. Conclusions

In this work, we propose a defense system called ShallowDeepNet (or G+D Net) that is able to resist in confronting adversarial attacks. The proposed system includes a shallow neural network and a deep neural network. The shallow neural network is responsible for executing the data preprocessing tasks, while the deep neural network (known as RazorNet) needs to perform the main data processing. The data preprocessing is defined as the generation of the adversarial examples (or error-contained data). It is done through fooling a shallow neural network coupled with an attacking engine that includes certain well-known attacks, namely FGSM, JSMA, DeepFool, C&W, and PGD. The generated adversarial examples from this engine are used in retraining the pretrained RazorNet. Inclusion of multiple elements into our system, namely detector neural network, transfer learning, adversarial training, and noise training, makes this system strong and robust enough to recognize and detect

clean and malicious data from different domains. The simulation results from running the biometric (fingerprint and iris) and the pill image data on this system proves its capability in detecting the malicious versions of these data with accuracies of 93.4% and 98.20% respectively.

Author Contributions: S.T. contributed in coming up with the ideas, running the experiments, and writing the manuscript. M.S. contributed in a detailed discussion of the project ideas and results. J.-S.Y. provided technical feedback and reviewed the manuscript. All authors read and confirmed the final manuscript.

Funding: There is no funding resource for this project.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing properties of neural networks. *arXiv* **2013**, arXiv:1312.6199.
2. Bastani, O.; Ioannou, Y.; Lampropoulos, L.; Vytiniotis, D.; Nori, A.; Criminisi, A. Measuring neural net robustness with constraints. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 2613–2621.
3. Gu, S.; Rigazio, L. Towards deep neural network architectures robust to adversarial examples. *arXiv* **2014**, arXiv:1412.5068.
4. Huang, R.; Xu, B.; Schuurmans, D.; Szepesvári, C. Learning with a strong adversary. *arXiv* **2015**, arXiv:1511.03034.
5. Jin, J.; Dundar, A.; Culurciello, E. Robust convolutional neural networks under adversarial noise. *arXiv* **2015**, arXiv:1511.06306.
6. Papernot, N.; McDaniel, P.; Wu, X.; Jha, S.; Swami, A. Distillation as a defense to adversarial perturbations against deep neural networks. In Proceedings of the 2016 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 22–26 May 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 582–597.
7. Rozsa, A.; Rudd, E.M.; Boult, T.E. Adversarial diversity and hard positive generation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 25–32.
8. Shaham, U.; Yamada, Y.; Negahban, S. Understanding adversarial training: Increasing local stability of supervised models through robust optimization. *Neurocomputing* **2018**, *307*, 195–204. [CrossRef]
9. Zheng, S.; Song, Y.; Leung, T.; Goodfellow, I. Improving the robustness of deep neural networks via stability training. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 4480–4488.
10. Mirjalili, V.; Ross, A. October. Soft biometric privacy: Retaining biometric utility of face images while perturbing gender. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*; IEEE: Piscataway, NJ, USA, 2017; pp. 564–573.
11. Jia, R.; Liang, P. Adversarial examples for evaluating reading comprehension systems. *arXiv* **2017**, arXiv:1707.07328.
12. Belinkov, Y.; Bisk, Y. Synthetic and natural noise both break neural machine translation. *arXiv* **2017**, arXiv:1711.02173.
13. Samanta, S.; Mehta, S. Towards crafting text adversarial samples. *arXiv* **2017**, arXiv:1707.02812.
14. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in neural information processing systems, Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
15. Tramèr, F.; Kurakin, A.; Papernot, N.; Goodfellow, I.; Boneh, D.; McDaniel, P. Ensemble adversarial training: Attacks and defenses. *arXiv* **2017**, arXiv:1705.07204.
16. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv* **2017**, arXiv:1706.06083.
17. Papernot, N.; McDaniel, P.; Goodfellow, I.; Jha, S.; Celik, Z.B.; Swami, A. Practical black-box attacks against machine learning. In Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, New York, NY, USA, 2–6 April 2017; pp. 506–519.
18. Połap, D.; Woźniak, M.; Wei, W.; Damaševičius, R. Multi-threaded learning control mechanism for neural networks. *Future Gener. Comput. Syst.* **2018**, *87*, 16–34.

19. Yu, L.; Zhang, W.; Wang, J.; Yu, Y. Seqgan: Sequence generative adversarial nets with policy gradient. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
20. Narayanan, B.N.; Hardie, R.C.; Balster, E.J. Multiframe Adaptive Wiener Filter Super-Resolution with JPEG2000-Compressed Images. Available online: https://link.springer.com/article/10.1186/1687-6180-2014-55 (accessed on 19 July 2019).
21. Narayanan, B.N.; Hardie, R.C.; Kebede, T.M. Performance analysis of a computer-aided detection system for lung nodules in CT at different slice thicknesses. *J. Med. Imaging* **2018**, *5*, 014504. [CrossRef] [PubMed]
22. Chivukula, A.S.; Liu, W. Adversarial Deep Learning Models with Multiple Adversaries. *IEEE Trans. Knowl. Data Eng.* **2018**, *31*, 1066–1079. [CrossRef]
23. Kwon, H.; Kim, Y.; Park, K.W.; Yoon, H.; Choi, D. Multi-targeted adversarial example in evasion attack on deep neural network. *IEEE Access* **2018**, *6*, 46084–46096. [CrossRef]
24. Shen, J.; Zhu, X.; Ma, D. TensorClog: An Imperceptible Poisoning Attack on Deep Neural Network Applications. *IEEE Access* **2019**, *7*, 41498–41506. [CrossRef]
25. Kulikajevas, A.; Maskeliūnas, R.; Damaševičius, R.; Misra, S. Reconstruction of 3D Object Shape Using Hybrid Modular Neural Network Architecture Trained on 3D Models from ShapeNetCore Dataset. *Sensors* **2019**, *19*, 1553. [CrossRef] [PubMed]
26. Carrara, F.; Falchi, F.; Caldelli, R.; Amato, G.; Becarelli, R. Adversarial image detection in deep neural networks. *Multimed. Tools Appl.* **2019**, *78*, 2815–2835. [CrossRef]
27. Li, Y.; Wang, Y. Defense Against Adversarial Attacks in Deep Learning. *Appl. Sci.* **2019**, *9*, 76. [CrossRef]
28. Grosse, K.; Manoharan, P.; Papernot, N.; Backes, M.; McDaniel, P. On the (statistical) detection of adversarial examples. *arXiv* **2017**, arXiv:1702.06280.
29. Gong, Z.; Wang, W.; Ku, W.S. Adversarial and clean data are not twins. *arXiv* **2017**, arXiv:1704.04960.
30. Metzen, J.H.; Genewein, T.; Fischer, V.; Bischoff, B. On detecting adversarial perturbations. *arXiv* **2017**, arXiv:1702.04267.
31. Feinman, R.; Curtin, R.R.; Shintre, S.; Gardner, A.B. Detecting adversarial samples from artifacts. *arXiv* **2017**, arXiv:1703.00410.
32. Carlini, N.; Wagner, D. Towards evaluating the robustness of neural networks. In Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 22–24 May 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 39–57.
33. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; Lempitsky, V. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* **2016**, *17*, 1–35.
34. Marchisio, A.; Nanfa, G.; Khalid, F.; Hanif, M.A.; Martina, M.; Shafique, M. CapsAttacks: Robust and Imperceptible Adversarial Attacks on Capsule Networks. *arXiv* **2019**, arXiv:1901.09878.
35. Xu, W.; Evans, D.; Qi, Y. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv* **2017**, arXiv:1704.01155.
36. Kannan, H.; Kurakin, A.; Goodfellow, I. Adversarial logit pairing. *arXiv* **2018**, arXiv:1803.06373.
37. Mopuri, K.R.; Babu, R.V. Gray-box Adversarial Training. *arXiv* **2018**, arXiv:1808.01753.
38. Neelakantan, A.; Vilnis, L.; Le, Q.V.; Sutskever, I.; Kaiser, L.; Kurach, K.; Martens, J. Adding gradient noise improves learning for very deep networks. *arXiv* **2015**, arXiv:1511.06807.
39. Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F.; Wattenberg, M. Smoothgrad: Removing noise by adding noise. *arXiv* **2017**, arXiv:1706.03825.
40. Gao, F.; Wu, T.; Li, J.; Zheng, B.; Ruan, L.; Shang, D.; Patel, B. SD-CNN: A shallow-deep CNN for improved breast cancer diagnosis. *Comput. Med. Imaging Graph.* **2018**, *70*, 53–62. [CrossRef] [PubMed]
41. Ernst, D.; Das, S.; Lee, S.; Blaauw, D.; Austin, T.; Mudge, T.; Kim, N.S.; Flautner, K. Razor: Circuit-level correction of timing errors for low-power operation. *IEEE Micro* **2004**, *24*, 10–20. [CrossRef]
42. Basu, S.; Karki, M.; Ganguly, S.; DiBiano, R.; Mukhopadhyay, S.; Gayaka, S.; Kannan, R.; Nemani, R. Learning sparse feature representations using probabilistic quadtrees and deep belief nets. *Neural Process. Lett.* **2017**, *45*, 855–867. [CrossRef]
43. CASIA-FingerprintV5. 2010. Available online: http://biometrics.idealtest.org/dbDetailForUser.do?id=7 (accessed on 26 December 2017).
44. CASIA-IrisV4. 2010. Available online: http://biometrics.idealtest.org/dbDetailForUser.do?id=4 (accessed on 26 December 2017).

45. 1k Pharmaceutical Pill Image Dataset. Available online: https://www.kaggle.com/trumedicines/1k-pharmaceutical-pill-image-dataset (accessed on 2 June 2018).
46. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2009**, *22*, 1345–1359. [CrossRef]
47. Bottou, L. Large-scale machine learning with stochastic gradient descent. In *Compstat'2010*; Physica-Verlag HD: Heidelberg, Germany, 2010; pp. 177–186.
48. Liu, X.; Cheng, M.; Zhang, H.; Hsieh, C.J. Towards robust neural networks via random self-ensemble. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 369–385.
49. Ranjan, R.; Sankaranarayanan, S.; Castillo, C.D.; Chellappa, R. Improving network robustness against adversarial attacks with compact convolution. *arXiv* **2017**, arXiv:1712.00699.
50. Song, Y.; Kim, T.; Nowozin, S.; Ermon, S.; Kushman, N. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. *arXiv* **2017**, arXiv:1710.10766.
51. Dai, X.; Gong, S.; Zhong, S.; Bao, Z. Bilinear CNN Model for Fine-Grained Classification Based on Subcategory-Similarity Measurement. *Appl. Sci.* **2019**, *9*, 301. [CrossRef]
52. Varior, R.R.; Haloi, M.; Wang, G. Gated siamese convolutional neural network architecture for human re-identification. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; Springer: Cham, Switzerland, 2016; pp. 791–808.
53. Ranjan, R.; Patel, V.M.; Chellappa, R. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *41*, 121–135. [CrossRef] [PubMed]
54. Twinanda, A.P.; Shehata, S.; Mutter, D.; Marescaux, J.; De Mathelin, M.; Padoy, N. Endonet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE Trans. Med. Imaging* **2016**, *36*, 86–97. [CrossRef] [PubMed]