Ad Hoc Networks 122 (2021) 102627

Contents lists available at ScienceDirect

Ad Hoc Networks

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

# Do not get fooled: Defense against the one-pixel attack to protect IoT-enabled Deep Learning systems☆

Muhammad Akbar Husnoo, Adnan Anwar *

*Centre for Cyber Security Research and Innovation (CSRI), School of Information Technology, Deakin University, Geelong, VIC 3216, Australia*

ARTICLE INFO

*Keywords:*
Internet of Things (IoT)
Industrial Internet of Things (IIoT)
One-pixel attack
Machine learning
Deep Learning
Adversarial machine learning
Attacks
Data integrity
Defenses
Image recovery

ABSTRACT

With the rapid evolution and proliferation of *Artificial Intelligence* (AI) throughout the past decade, industry-wide Deep Learning (DL) applications within IIoT ecosystems are expected to rise exponentially within this coming decade. However, recent studies have highlighted the extreme sensitivity and vulnerability of modern Deep Neural Networks (DNNs) to adversarial examples. While being imperceptible and benign to the naked human eye, those adversarial perturbations crafted from *one-pixel attacks* can drastically impact the accuracy of DNNs. The consequences of those adversarial attacks in critical applications such as healthcare and autonomous vehicles can prove disastrous. Therefore, the urge to ensure the robustness and resilience of DL systems against one-pixel attacks has recently started attracting stupendous attention from both academicians and industry professionals. Throughout this paper, we first develop a one-pixel threat model within an IIoT ecosystem and propose a novel image recovery defense mechanism to detect and mitigate one-pixel attacks based on Accelerated Proximal Gradient approach. Experimental results proved that our proposed solution is highly effective and efficient in detecting and mitigating one-pixel attacks in state-of-the-art neural networks, more specifically LeNet and ResNet, using the CIFAR10 and MNIST datasets.

## 1. Introduction

Contemporary advancements in technology and wireless communications have given rise to one of the hottest concepts of the 21st century termed as Internet of Things (IoT) [1]. Accompanied by a vast array of technological [2] and business benefits [3], IoT has poised to crossover several critical application domains including healthcare, transport, energy as well as industry sectors [4]. McKinsey [5] predicts a sharp growth of IoT usage within the next couple of years, reaching 43 billion connected devices by 2023. As the number of internet-connected devices is growing exponentially, enormous amounts of data are being collected and generated by those devices. In the case of critical infrastructure, the taming of the data deluge is of high priority as efficacy is strictly related to the real-time processing of big data which is gathered via several edge IoT-connected devices [6].

As aforementioned, the underlying automation and intelligence behind IoT paradigm lies on the shoulders of real-time efficient and effective big data analytics [7]. In parallel with the rise of IoT, another interesting computing paradigm known as Deep Learning (DL) [8] has gained momentum during the past decade. The past few years have overseen the amalgamation of mature DL technologies within the IoT ecosystem to solve the complicated task of deriving information and inferences which were previously unaccomplished by other conventional paradigms [9]. Having shown previous remarkable results in various fields [10] such as games development [11], object [12] and speech recognition [13], natural language modeling and processing [14], and virtual assistance [15], it is anticipated that DL will play a pivotal role in enabling smarter critical Industrial IoT (IIoT) infrastructure [9].

However, the array of benefits brought about by the integration of those two state-of-the art computing paradigms are short-lived as adversaries are being lured to explore their extreme vulnerabilities [16]. The vast amount of sensitive data being collected by IoT-enabled critical infrastructure which is then being processed using DL technologies is attracting more and more black hats for political, commercial and monetary gains [17]. In particular, during the past couple of years, the number of adversarial attacks on DL networks has drastically increased [18]. A number of researches [19–27] in both academia and industry have highlighted a stupendous prevailing security threat among existing DL technologies. Deep Neural Networks (DNNs) can easily be fooled through simple perturbations of the data which is imperceptible to the naked human eye.

☆ This document is the results of the research project funded by the Centre for Cyber Security Research and Innovation (CSRI), School of Information Technology, Deakin University, Australia.

* Corresponding author.
*E-mail addresses:* mahusnoo@deakin.edu.au (M.A. Husnoo), adnan.anwar@deakin.edu.au (A. Anwar).

https://doi.org/10.1016/j.adhoc.2021.102627
Received 1 March 2021; Received in revised form 3 June 2021; Accepted 20 July 2021
Available online 28 July 2021

This extreme vulnerability of DNNs to subtle imperceptible adversarial perturbations has now become a major concern which hinders its acceptance and application to IoT-enabled critical infrastructure [28]. To mitigate this major issue, researchers from all over the globe are actively involved in devising and formulating effective defense strategies to increase resiliency of DNNs against adversarial attacks which has given rise to one of the hottest topics of research known ad Adversarial Deep Learning (AdvDL) [29]. In particular, Su et al. [30] proposed one-pixel attack, an extreme case of adversarial perturbation through the change of only one single pixel that successfully fooled neural networks. In this view, this paper aims at increasing the robustness and resilience of DNNs by developing an effective defense strategy to both detect and mitigate one-pixel attacks on state-of-the-art Convolutional Neural Networks (CNNs) used in an IoT ecosystem. The high-level idea of our proposed algorithm is to achieve the effect of passive defense against one-pixel attacks without the need to train another model by recovering the malicious pixel in the adversarial sample and thus, reconstructing the original image. In summary, the main contributions of our paper are listed below:

- Develop and demonstrate the threat model of one-pixel attacks in an IIoT ecosystem.
- Propose an image recovery and reconstruction approach for defending state-of-the-art DNNs against one-pixel attacks.
- Experimentally validate the effectiveness of our proposed algorithm by achieving a high accuracy against one-pixel attacks. Results shows that it is possible to recover the original low-rank images with high accuracy.

The remainder of this paper is organized as follows: in Section 2, we discuss the defense mechanisms against one-pixel attacks along with their shortcomings. In Section 3, we discuss the threat model of one-pixel attacks in an IIoT ecosystem as well as a detailed overview of one-pixel attacks. In Section 4, we propose our image recovery algorithm that can eliminate one-pixel attacks while in Section 5, we highlight the experimental configurations and set-ups required. In 6, we discuss and compare the experimental results of our proposed solution. Finally, we give some future research directions of our proposed approach and conclude our work in Section 7.

## 2. Related works

### 2.1. Existing attacks on DNNs

As aforementioned, DNNs have demonstrated impressive capabilities in several application domains such as games development [11], object [12] and speech recognition [13], natural language modeling and processing [14], and virtual assistance [15]. DNNs have proved to even exceed humans at certain tasks [10] which has increased their adoption to several real-world critical scenarios. However, recent literature [19–27] have exposed the extreme vulnerability of DNNs to subtle and imperceptible adversarial perturbations. Szedegy et al. [19] first pioneered adversarial attacks on DNNs by generating adversarial samples through searching the minimal distorted adversarial $x'$ using Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) algorithm. As L-BFGS attack was found to be computationally expensive, Goodfellow et al. [21] proposed Fast Gradient Sign Method as a one-step gradient update along the descending direction of the sign of gradient for each pixel to generate adversarial examples.

Similarly, several other attacks including [20,22,23,25–27] have been proven successful in highlighting the extreme susceptibility of DNNs to adversarial perturbations. More recently, Moosavi-Dezfooli et al. [24] devised a new attack known as DeepFool to find the shortest path such that $x$ can go beyond the decision boundary and hence generate adversarial samples. Similarly, the authors in [27] proposed a novel type of attacks known as Universal Adversarial Perturbations (UAP) to find the single optimal quasi-imperceptible fixed image-agnostic perturbation that enables the misclassification of a DNN on all test samples. From those two discussed attacks, it can be seen that research trends have shifted to finding the minimal perturbation required to force misclassification with high accuracy in state-of-the-art deep neural networks.

### 2.2. One-pixel attacks

However, in contrast to UAP which modifies several pixels, Su et al. [30] proposed one-pixel attack in 2017 which is an extreme interesting case that reveals the intriguing cons of DL technologies. Through the alteration of a single pixel of an image which is indeed imperceptible to the human eye, one-pixel attack surprisingly causes state-of-the-art CNNs to wrongly classify images with high accuracy. It is a gray-box attack that requires very limited information i.e. only the confidence values of labels output by DNNs. The generation of such adversarial examples is an optimization problem which can be mathematically denoted as:

$$\begin{aligned} &\max_{e(x)^*} \ f_{adv}(x + e(x)) \\ &s.t. \ \|e(x)\|_0 \leq d \end{aligned} \tag{1}$$

where $d = 1$ in case of one-pixel attack. In view of finding the optimal solution, the researchers utilized a population based evolutionary optimization algorithm as it requires no gradient information for optimization and therefore, the objective function does not need to be differentiable or previously known. In more succinct details, the researchers used Differential Evolution algorithm to search for the optimal pixel that crafts an adversarial sample by changing the optimal pixel value. Having evaluated their attack on ImageNet and CIFAR-10 datasets, it was found that 67.97% of the CIFAR-10 data and 16.04% can be attacked by changing the value of only one pixel which vastly exposes the vulnerability of DNNs to attacks of low dimensionality.

### 2.3. Existing solutions against one-pixel attacks

Throughout recent literature, we identified three proposed solutions against one-pixel attacks. Chen et al. [31] proposed Patch Selection Denoiser that *removes few of the potential attacking pixels* in the whole image. However, it is important to note that their approach has received less acclaim due to the degradation of the image through the denoising neural network even though the authors claimed to have achieved a successful defense rate of 98.6% against one-pixel attacks. Similarly, Liu et al. [32] proposed a three step image reconstruction algorithm which *removes few of the potential attacking pixels* in the whole image. However, the authors also mentioned that their parameters were not optimal and their approach may lead to information loss in certain scenarios. Lastly, Shah et al. [33] proposed the usage of an Adversarial Detection Network (ADNet) for detection of one-pixel attacks. The drawbacks of this method includes the *complete discrimination and rejection of the adversarial input*, the training of a new network which is computationally expensive and the accuracy of the solution proposed is lower for the detection of one-pixel attack. Table 1 compares the three existing methodologies

Following the review of those previously proposed defense strategies against one-pixel attacks as highlighted in Table 1, we believe that there still is a need for a better defense strategy that preserves the integrity of the image data. Therefore, in this manuscript, we propose a more effective algorithm that estimates the full recovery of the malicious pixel injected by the attacker and reconstructs the original image with high accuracy.

## 3. Threat model of one-pixel attack in IIoT ecosystem

As earlier mentioned, Su et al. [30] proposed an extreme case of adversarial attacks known as one-pixel attack through the perturbation

**Table 1**
Contrast against existing defenses.

| Work | Year | Methodology | Technique | Dataset used |
|---|---|---|---|---|
| [31] | 2019 | A SRResNet is used for image denoising. The denoised and original images are then divided into three channels (RGB). A 3X3 patch will scan through the attacked image and corresponding denoised image. By comparing the absolute difference between the two patches, the patch from the original image replaces that of the original image. | Noise2Noise model & Patch Denoising | CIFAR-10 |
| [32] | 2020 | A 3-stage image reconstruction algorithm is proposed. The first stage involves the construction of a 32X32 difference map to record the extent to which each pixel differs based on its neighboring ones. The second step includes the generation of a 32X32X3 average map to store the average RGB values based on the difference map. Lastly, the each pixel is scanned sequentially in an attempt to remove the attacking pixel by comparison against the pre-defined threshold. | Image Reconstruction | CIFAR-10 |
| [33] | 2020 | Using a ResNet-152 which is made up several convolutional and pooling layers, the authors trained an Adversarial Detection Network (ADNet) which enables the detection and rejection of adversarial samples prior to feeding them to a classifier. | Adversarial Sample Discrimination | CIFAR-10, CIFAR-100 & MNIST |
| [Our Work] | 2021 | We propose an recovery algorithm based on Accelerated Proximal Gradient approach to detect and recover the attacked pixel. | Image Recovery | CIFAR-10 & MNIST |

of a single pixel. While the initially proposed attack is a stand-alone one, we believe that this attack can have severe implications if extended to an IIoT ecosystem. While edge computing paradigm has proven to solve several data traffic issues that were previously being encountered in IIoT networks [34], it opens up new vulnerability points which can be easily leveraged by adversaries. That being said, in Fig. 1 below, it can be seen that an adversary can easily gain access to the image data collected by edge sensor devices. For the purpose of this work, we consider the main threat model to be adversarial perturbations at the edge. The potential attacker can perturb the image data during transmission to the data collection device. This is due to the increasing variety and complexity of edge devices connected to an IIoT network which makes it hard to secure the data at the edge [35]. Furthermore, injecting adversarial samples at the edge devices prior to data aggregation can be considered as a gray-box attack where the adversary requires limited knowledge of the internal IIoT network. The aggregated data is then transmitted to a cloud storage after which the cloud analytics platform processes the real-time data into inferences. However, in the case of an adversarial one-pixel attack, the adversarially perturbed samples are fed to DL models and forces misclassification. This leads to negative inferences which may fully disrupt the whole IIoT ecosystem.

However, in the real-world scenario, this is not only the case. Therefore, we depict the other potential threats as in Fig. 1. From Fig. 1, it can be seen that adversarial one-pixel perturbations can occur at the edge devices/data aggregators/cloud platforms, during transmission from the edge to the data aggregators, during transmission from data aggregators to cloud server and thereon to the cloud analytics platform.

The generation of such adversarial examples is an optimization problem which can be mathematically denoted as:

$$\begin{aligned} &\max_{e(x)^*} \ f_{adv}(x + e(x)) \\ &s.t. \ \|e(x)\|_0 \leq d \end{aligned} \tag{2}$$

where $d = 1$ in case of one-pixel attack. In view of finding the optimal solution, the researchers utilized a population based evolutionary optimization algorithm as it requires no gradient information for optimization and therefore, the objective function does not need to be differentiable or previously known. Having evaluated their attack on ImageNet and CIFAR-10 datasets, it was found that 67.97% of the CIFAR-10 data and 16.04% can be attacked by changing the value of only one pixel which vastly exposes the vulnerability of DNNs to attacks of low dimensionality [30].

## 4. Proposed defense mechanism

*Principal Component Analysis* (PCA) is one of the oldest and most popular statistical technique to reduce the high-dimensionality of large datasets which preserving as much statistical information as possible [36]. However, one of the major drawbacks of classical PCA is its extreme sensitivity to grossly corrupted or outlying observations. While in many cases the outliers are removed before performing PCA, gross errors tend to be ubiquitous due to arbitrarily corrupted measurements [37]. Therefore, Wright et al. [37] proposed *Robust Principal Component Analysis* (RPCA) as a modification of the classical PCA that works well with grossly corrupted or outlying observations. The basic mathematical formulation behind RPCA relies on:

$$D = A + E \tag{3}$$

where $D \in \mathbb{R}^{m \times n}$. is the observed data matrix obtained by corrupting some entries of a low-rank matrix, $A \in \mathbb{R}^{m \times n}$ with an additive error, $E \in \mathbb{R}^{m \times n}$. For RPCA, given $D = A + E$, $A$ and $E$ are unknown but $A$ is known to be low-rank and $E$ is known to be sparse. That being said, A can then be recovered through the conceptual solution of probing for the lowest-ranked $A$ that generated $D$ s.t. the constraint that the error, $E$ is sparse.

### 4.1. Problem formulation

Suppose the given adversarial image, $x'$ is a large matrix where $x' \in \mathbb{R}^{m \times n}$. $x'$ is the result of adversarially perturbing the original image, $x$ with a perturbation $\delta$ where $x$ is also a matrix and $x \in \mathbb{R}^{m \times n}$. Therefore, the author formulates this given problem as follows:

$$x' = x + \delta \tag{4}$$

In this mathematical setting, the proposed solution seeks to find the optimal estimate of the low-rank matrix, $x$ which translates to the following convex optimization problem:

$$\begin{aligned} &\min_{x,\delta} \|x\|_* + \lambda |\delta|_1 \\ &s.t. \ x' = x + \delta \end{aligned} \tag{5}$$

where $\|x\|_*$ is the nuclear norm of matrix $x$, $|\delta|_1$ is the sum of the absolute values of $\delta$ and $\lambda$ is a positive weighting parameter. This optimization is commonly referred to *Robust Principal Component Analysis* (RPCA) [37] which allows to recover the low-rank original image, $x$ even in the presence of gross errors (in this case, the attacked pixel acts as the outlying observation).

The same problem formulation can be applied to other IoT-enabled data integrity attacks that are similar to one-pixel attacks within an IoT-enabled Deep Learning system. In other scenarios where attackers are injecting manipulated information into the streaming data, the problem can be formulated using similar approaches as shown in Eq. (4). In such

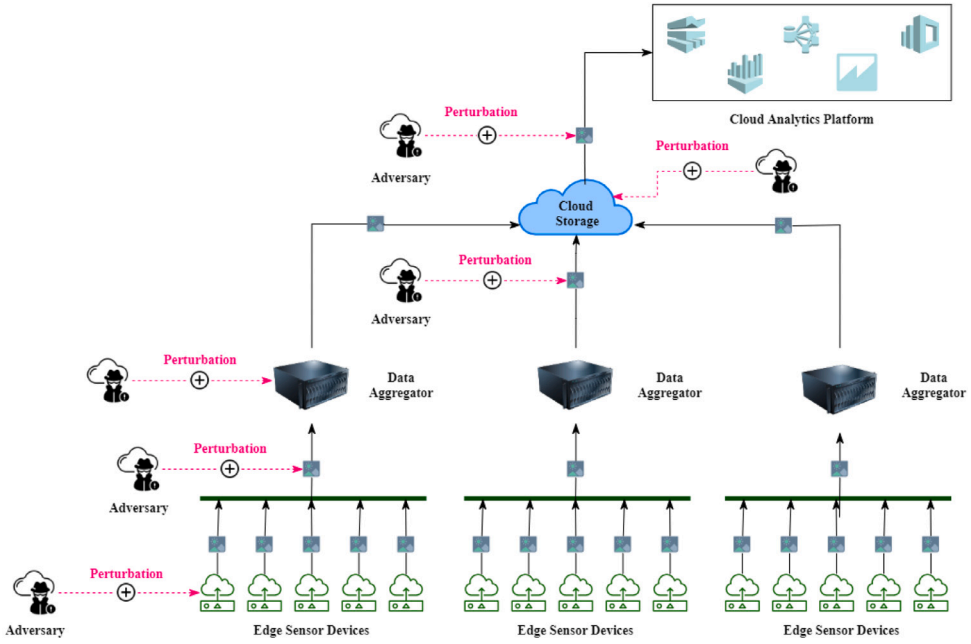



**Fig. 1.** Threat Model of one-pixel attack in an IIoT Ecosystem..

scenarios, if the measurement vector obtained from sensors is $z$ and the attack vector is $a$, then the corrupted measurement signal will be:

$$z_a = z + a \tag{6}$$

where $z$ is the original sensor information from IoT devices and $a$ is the attack vector. The above formulation in Eq. (6) has similarity to that of the one-pixel attack model as formulated in Eq. (4). Hence, the proposed problem formulation can also be applied for other IoT-enabled systems which face similar data integrity threats on sensor information. However, to limit the focus of this paper, we have scoped the work on one-pixel attacks.

### 4.2. Proposed solution

In this section, we will provide a thorough step-by-step explanation of the proposed strategy as in the following. A more visual representation of our approach can be found in the below block diagram in Fig. 2.

**Step 1**: In the first step, we read the adversarial image, $x'$ into a matrix format. The length of the matrix is determined. If $x'$ is a three dimensional matrix, we break down the matrix into 3 different 2-D matrices.

**Step 2**: During the second step, sparse matrix separation technique is utilized to separate gross error (adversarial pixel, $x'$) from the 2-D matrices from Step 1. Please refer to Section 4.1 for more details about the Problem Formulation. After formulating this adversarial attack to a RPCA optimization problem in Section 4.1, it can be observed that Eq. (5) is comes from a more generic class of optimizations known as proximal gradient algorithms [38,39] such that:

$$\begin{aligned} &\min_{X\in\mathcal{H}} F(X) = \mu\, g(X) + F(X) \\ &s.t.\ \mathcal{A}(X) = b \end{aligned} \tag{7}$$

where $\mathcal{H}$ is the real Hilbert space, $g(X)$ is a continuous convex function, $F(X)$ is the penalty function and $\mu$ is the relaxation parameter whereby $\mu > 0$.

Therefore, an Accelerated Proximal Gradient (APG) based technique previously proposed by Lin et al. [40] can be used to estimate and recover the original 2-D matrix $x$ which is free from sparse adversarial perturbations. If the penalty function, $F(X)$, is linearized by a separable quadratic approximation, $Q(X, Y)$, then the aforementioned APG based technique minimized the approximations of $Q(X, Y)$ instead of $F(X)$ for any chosen point Y as in the following:

$$Q(X,Y) = f(Y) + \langle \nabla f(Y), X - Y\rangle + \frac{L_f}{2}\|X - Y\|^2 + \mu g(X) \tag{8}$$

where $L_f$ is the Lipschitz constant that allows the monitoring and achievement of a better optimization efficiency. Given $G = Y - \frac{1}{L_f}\nabla f(Y)$, the objective function becomes as in the following:

$$arg\ min_x \quad Q(X,Y) = arg\ min_x \quad \mu g(X) + \frac{L_f}{2}\|X - G\|^2 \tag{9}$$

This RPCA problem is the solvable using the previously mentioned APG based technique as discussed in details in [40]. Our proposed strategy is developed as in Algorithm 1.

**Algorithm 1:** Proposed Defense mechanism to detect and mitigate one-pixel attacks.

**Input** : $x' \in \mathbb{R}^{m\times n}$

Initialize Sequence Parameter, $t_k$, Tuning Parameter, $\lambda$ & Relaxation Parameter, $\mu_0$.

**while** *No Convergence* **do**

$\mathbf{Y^x_k} \longleftarrow \mathbf{x_k} + \frac{\mathbf{t_{k-1}-1}}{\mathbf{t_k}}(\mathbf{x_k} - \mathbf{x_{k-1}})$

$\mathbf{Y^\delta_k} \longleftarrow \delta_\mathbf{k} + \frac{\mathbf{t_{k-1}-1}}{\mathbf{t_k}}(\delta_\mathbf{k} - \delta_\mathbf{k-1})$;

$\mathbf{G^A_k} \longleftarrow \mathbf{Y^x_k} - \frac{\mathbf{1}}{\mathbf{L_f}}(\mathbf{Y^x_k} + \mathbf{Y^\delta_k} - \tilde{\mathbf{x}})$ ;

$\mathbf{G^\delta_k} \longleftarrow \mathbf{Y^\delta_k} - \frac{\mathbf{1}}{\mathbf{L_f}}(\mathbf{Y^x_k} + \mathbf{Y^\delta_k} - \tilde{\mathbf{x}})$;

$[\mathbf{U}, \mathbf{S}, \mathbf{V}] = \mathbf{svd}(\mathbf{G^A_k})$;

$\mathbf{x_{k+1}} = \mathbf{U}\xi_{\frac{\mu_\mathbf{k}}{\mathbf{L_f}}}[\mathbf{S}]\mathbf{V^T}$ and ;

$\delta_{\mathbf{k+1}} = \xi_{\frac{\lambda\mu_\mathbf{k}}{\mathbf{L_f}}}[\mathbf{G^\delta_k}]$;

$t_{k+1} \longleftarrow \frac{1+\sqrt{4t_k^2+1}}{L_f}$ and ;

$\mu_{k+1} \longleftarrow \max(\eta\mu_k, \bar{\mu})$ ;

$k \longleftarrow k + 1$ ;

**end**

**Output**: $\mathbf{x} \longleftarrow \mathbf{x_k}, \delta \longleftarrow \delta_\mathbf{k}$

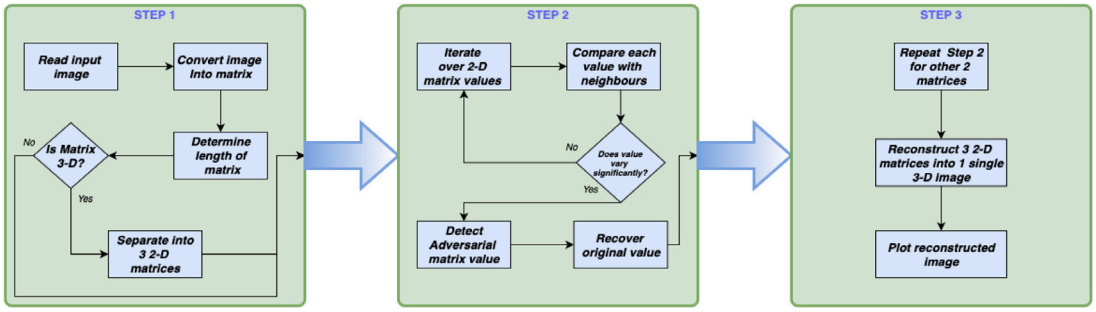



**Fig. 2.** Block Diagram representing the flow of our proposed solution.

**Step 3**: Finally, the recovered three 2-D matrices of the adversarial image, $x'$ is are reconstructed to obtained an estimation of the original image, $x$.

The results obtained from the proposed defense mechanism are presented in the following Section 6

## 5. Experimental setup

### 5.1. Dataset

In order to train the DNNs and to perform the one-pixel attack on the images, we used two datasets namely CIFAR-10 [41] and GTSRB datasets [42]. CIFAR-10 dataset consists of 60000 RGB images with a 32 × 32 pixel resolution. 50000 images are for training and the remaining 10000 images are for testing. The data are of exactly 10 classes namely 'airplane', 'automobile', 'bird', 'cat', 'deer', 'dog', 'frog', 'horse', 'ship' and 'truck'. The dataset is broken down into 6 batches of 10000 images each whereby 5 batches are for training and 1 is for testing. The testing batch consists of exactly 1000 images that are randomly selected from each class.

On the other hand, GTSRB dataset consists of 51839 RGB images of 43 different classes of traffic signs with images from different light conditions and backgrounds. The dataset is split into 39209 training images and 12630 test data.

### 5.2. Deep neural networks

To successfully demonstrate the one-pixel attack and the efficacy of the proposed defense mechanism, we made use of two state-of-the-art DNNs namely LeNet and ResNet. LeNet [43] and ResNet [44] were chosen for chosen relatively simple, straightforward architecture to be implemented as well as fast training times.

### 5.3. Experimental setup

The authors firstly replicate and simulate the adversarial one-pixel attack using Google Colab and Python. The two aforementioned DNNs are first trained using CIFAR-10 and GTSRB dataset and saved. The pre-trained DNNs are then used to test the attack on a number of images and prove to be successful.

On the other hand, the proposed algorithm is implemented on MATLAB R2020b. The algorithm is run on a MacBook Air 2017 with 2.3 GHz dual-core Intel Core i5 and 8gb of RAM.

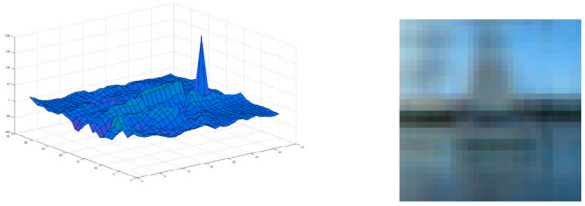
## 6. Results, analysis & discussion

### 6.1. Experiments with varying $\lambda$

As previously mentioned in Section 4, $\lambda$-parameter tuning is one of the most important tasks to ensure the effectiveness of the proposed algorithm. To demonstrate the change caused by the varying the value of lambda, the authors present the surface plots of the attack vector, $\delta$ at different $\lambda$-values on the adversarial images in Figs. 3 and 4:

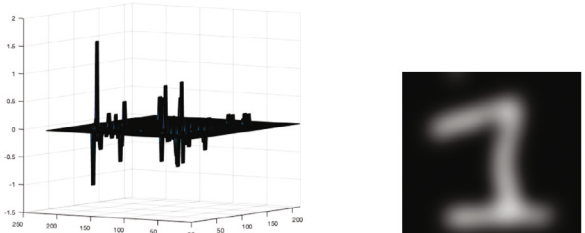
**Fig. 3.** 'Ship' image classified as 'Airplane' after one-pixel attack (CIFAR-10).

**Fig. 4.** '1' image classified as '4' after one-pixel attack (MNIST).



(a) Surface Plot of Attack Vector (b) Image After Recovery

**Fig. 5.** At $\lambda = 0.1$ (CIFAR-10)



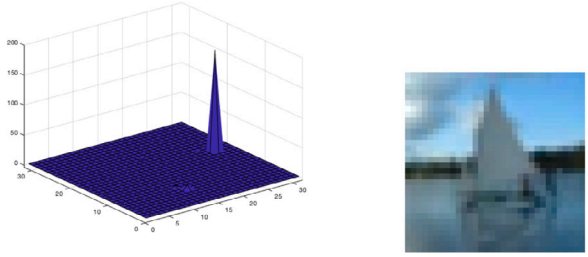
(a) Surface Plot of Attack Vector (b) Image After Recovery

**Fig. 6.** At $\lambda = 0.1$ (MNIST)

1. When $\lambda = 0.1$:
   At $\lambda = 0.1$, it can be seen that the algorithm detects several peeks and troughs apart from the highest peak which is the adversarially changed pixel (see Fig. 5).Therefore, the algorithm

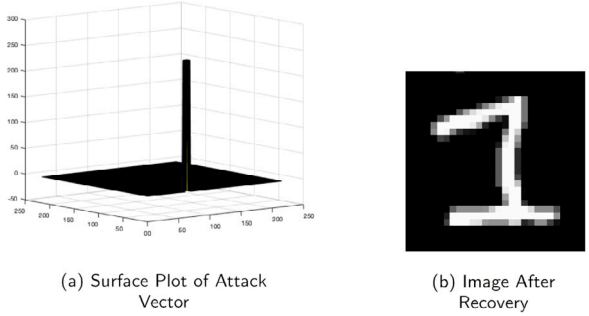
**Table 2**
Results Validation against previously existing solutions using similar datasets.

| Dataset | Work | Defense methodology | Defense technique | Accuracy |
|---|---|---|---|---|
| *CIFAR-10* | [31] | A SRResNet is used for image denoising. The denoised and original images are then divided into three channels (RGB). A 3X3 patch will scan through the attacked image and corresponding denoised image. By comparing the absolute difference between the two patches, the patch from the original image replaces that of the original image. | Noise2Noise model & Patch Denoising | 98.6% |
| | [32] | 3-stage image reconstruction algorithm is proposed. The first stage involves the construction of a 32X32 difference map to record the extent to which each pixel differs based on its neighboring ones. The second step includes the generation of a 32X32X3 average map to store the average RGB values based on the difference map. Lastly, the each pixel is scanned sequentially in an attempt to remove the attacking pixel by comparison against the pre-defined threshold. | Image Reconstruction | 91% |
| | [33] | Using a ResNet-152 which is made up several convolutional and pooling layers, the authors trained an Adversarial Detection Network (ADNet) which enables the detection and rejection of adversarial samples prior to feeding them to a classifier. | Adversarial Sample Discrimination | 98.7% |
| | [Our Proposed Solution] | We propose an recovery algorithm based on Accelerated Proximal Gradient approach to detect and recover the attacked pixel. | Image Recovery | 98.7% |
| *MNIST* | [33] | Using a ResNet-152 which is made up several convolutional and pooling layers, the authors trained an Adversarial Detection Network (ADNet) which enables the detection and rejection of adversarial samples prior to feeding them to a classifier. | Adversarial Sample Discrimination | 97.7% |
| | [Our Proposed Solution] | We propose an recovery algorithm based on Accelerated Proximal Gradient approach to detect and recover the attacked pixel. | Image Recovery | 98.2% |



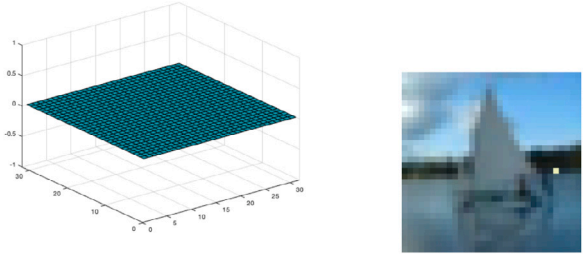
(a) Surface Plot of Attack Vector (b) Image After Recovery

**Fig. 7.** At $\lambda = 0.5$ (CIFAR-10)



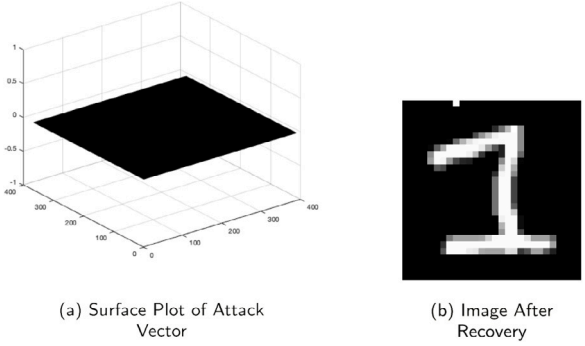
(a) Surface Plot of Attack Vector (b) Image After Recovery

**Fig. 8.** At $\lambda = 0.5$ (MNIST)



(a) Surface Plot of Attack Vector (b) Image After Recovery

**Fig. 9.** At $\lambda = 1$ (CIFAR-10)



(a) Surface Plot of Attack Vector (b) Image After Recovery

**Fig. 10.** At $\lambda = 1$ (GTSRB)

tends to recover the matrix such that other non-attacked pixels are changed. This results in poor performance and image degradation of the recovered image quality as shown in Fig. 6(b).

2. When $\lambda = 0.5$:

At $\lambda = 0.5$, it can be seen that the algorithm detects one extreme peak and few relatively benign peaks with the highest one being the adversarially modified pixel (see Fig. 7). Therefore, it can be concluded that $\lambda = 0.5$ is the most optimal value for the algorithm whereby there is very less amount of information

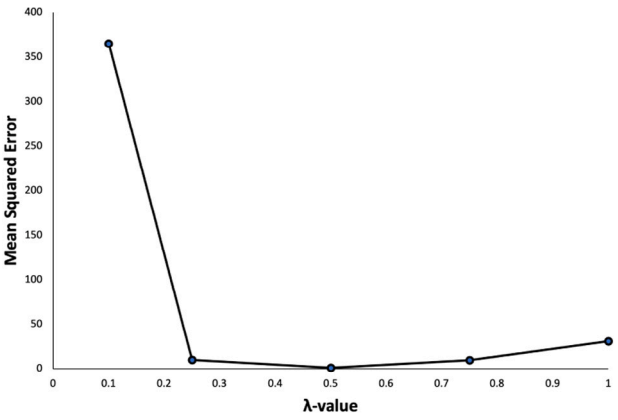



**Fig. 11.** Plot of Mean Squared Error vs $\lambda$.

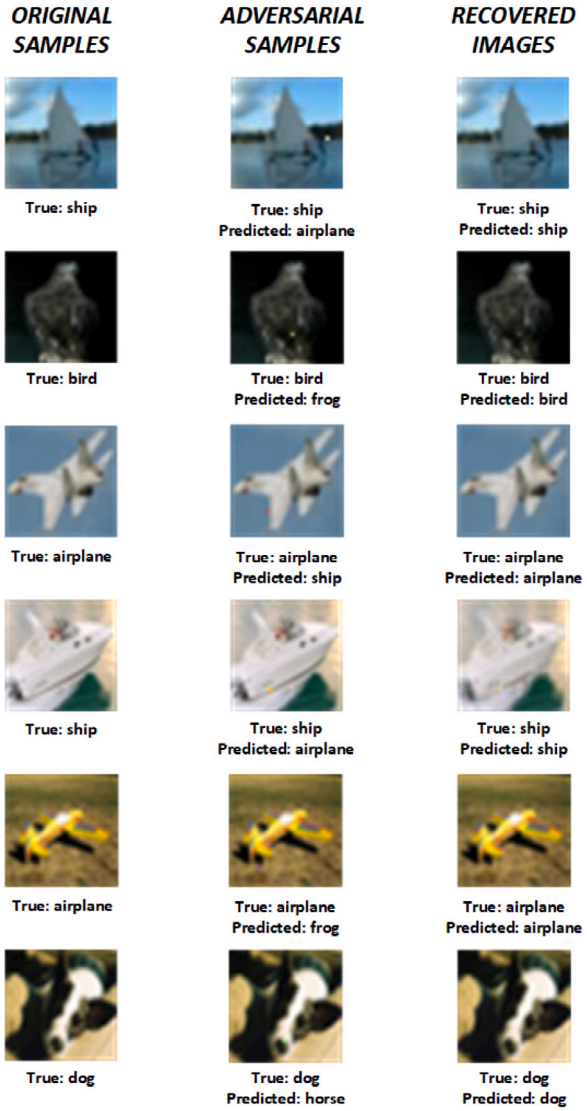



**Fig. 12.** Experimental Results of the proposed algorithm on CIFAR-10.

loss and the estimated image recovered is close to that of the original sample. As shown in Fig. 8 below, there are relatively no perceptible degradation between the image quality of Fig. 3 and Fig. 8(b).

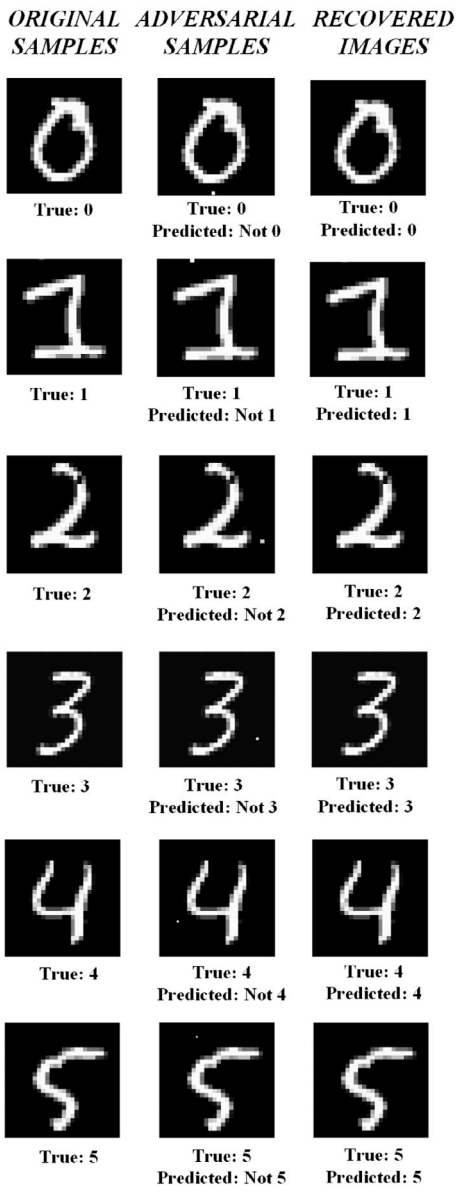



**Fig. 13.** Experimental Results of the proposed algorithm on MNIST.

3. When $\lambda = 1$:
   At $\lambda = 1$, it can be seen that the algorithm is unable to detect any adversarially changed pixel which results in the least optimal results. It can be seen from Fig. 10(b) and Fig. 9(b) that the adversarially changed pixel is still existing in both cases.

### 6.2. Choosing $\lambda$-value

In view of choosing the optimal $\lambda$-tuning parameter, we computed the Mean Squared Error (MSE) between the original image before attack and the recovered image. A plot of MSE vs $\lambda$ is as shown in Fig. 11. From the plot it can be seen that the MSE is the least at $\lambda = 0.5$ which is then chosen as the optimal $\lambda$-value.

### 6.3. Results & discussion

After having set up the experiments and chosen the most optimal lambda as shown in the previous Section 6.1, we validate the results

of the proposed image recovery algorithm on 6 different adversarial image as elaborated in the following:

Figs. 12 and 13 above show the experimental results obtained after testing the proposed image recovery solution on 6 different images from two benchmark datasets. The first columns of Figs. 12 and 13 represent original images directly from CIFAR-10 & MNIST dataset. The second columns of Figs. 12 and 13 represent the adversarial images crafted from one-pixel attack. It can be seen from the labels that the adversarially changed pixel results in misclassification of the images. However, after applying the proposed algorithm for recovering the images as shown in the third columns of Figs. 12 and 13, the DNNs are able to correctly classify the images.

Furthermore, to compute the accuracy of the algorithm, the author uses a simple formula as in the following:

$$Accuracy = \frac{Successful\ Mitigations}{Total\ of\ Images\ Tested} \times 100 \tag{10}$$

After testing the algorithm on images from both benchmark datasets, it was found that the proposed solution successfully mitigated one-pixel attacks on both benchmark datasets with high accuracy (see Table Table 2). That being said, the accuracy of our proposed method is therefore 98.7% on CIFAR-10 and 98.2% MNIST dataset. Therefore, it can be deduced that the proposed solution is highly effective at detecting and mitigating one-pixel attacks as well as recovering the images without perceptible information loss.

## 7. Conclusion

Throughout this manuscript, we have extended and investigated the fragility and vulnerability of DNNs to subtle and imperceptible adversarial one-pixel perturbations within an IIoT ecosystem. Following a review of the previously proposed solutions in existing literature, it was highlighted that they either degrade the quality of the image through adversarial pixel removal or to completely reject the adversarial sample. Therefore, there was still a dearth of an effective defense strategy to combat one-pixel attacks. In this view, we proposed and devised a better defense strategy based Accelerated Proximal Gradient [40] to estimate and recover the original image from an adversarial image. Experimental validation results in Fig. 12 revealed that our proposed solution is highly effective and efficient in detecting and mitigating one-pixel attacks in state-of-the-art neural networks, LeNet and ResNet, using the CIFAR10 & MNIST datasets.

Furthermore, as our research can motivate further exploration into adversarial attacks, we present some few future research directions which can improve our results further. In the view of recovering low rank matrices using RPCA, the proposed solution using Dual Partial Proximal Gradient approach [40] at $\lambda = 0.5$ proves to be effective in detecting and mitigating adversarial attacks. However, it is worth nothing that the image recovery estimation in some cases slightly degrades the quality of the original sample by detecting and recovering some *false adversarial pixels*. This is mostly due to the value of $\lambda$. Therefore, the author suggests to devise a strategy to find the optimal $\lambda$-value to decrease the information loss. Lastly, the proposed solution and algorithm is currently a stand-alone solution whereby the researcher has to first recover the image prior to feeding it to the DNN for classification. Therefore, the author proposes, as a future research direction, to concentrate on designing a DNN layer that incorporates the proposed algorithm as a pre-processing layer prior to classification. This layer can then be added to other state-of-the-art DNN architectures to improve robustness and resiliency.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References


[1] X. Cui, The internet of things, in: The Internet of Things, Palgrave Macmillan, UK, 2016, pp. 61–68, http://link.springer.com/10.1057/9781137505545_7.

[2] S. Li, L.D. Xu, S. Zhao, The internet of things: a survey, Inf. Syst. Front. 17 (2) (2015) 243–259.

[3] D. Newman, 5 iot trends to watch in 2021, 2020, https://www.forbes.com/sites/danielnewman/2020/11/25/5-iot-trends-to-watch-in-2021/. (Accessed 25 February 2021).

[4] M. Asplund, S. Nadjm-Tehrani, Attitudes and perceptions of iot security in critical societal services, IEEE Access 4 (2016) 2130–2138.

[5] D. Fredrik, P. Mark, R. Alexander, S. Jonathan, Growing opportunities in the internet of things, 2019, https://www.mckinsey.com/industries/private-equity-and-principal-investors/our-insights/growing-opportunities-in-the-internet-of-things. (Accessed 25 February 2021).

[6] R.A. Khalil, N. Saeed, M. Masood, Y.M. Fard, M.S. Alouini, T.Y. Al-Naffouri, Deep learning in the industrial internet of things: Potentials, challenges, and emerging applications, IEEE Internet Things J. (2021) 1.

[7] M. Marjani, F. Nasaruddin, A. Gani, A. Karim, I.A.T. Hashem, A. Siddiqa, I. Yaqoob, Big iot data analytics: Architecture, opportunities, and open research challenges, IEEE Access 5 (2017) 5247–5261.

[8] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.

[9] T.J. Saleem, M.A. Chishti, Deep learning for internet of things data analytics, Procedia Comput. Sci. 163 (2019) 381–390.

[10] L. Deng, A tutorial survey of architectures, algorithms, and applications for deep learning, APSIPA Trans. Signal Inf. Process. 3 (2014).

[11] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, M. Riedmiller, Playing atari with deep reinforcement learning, 2013, arXiv:1312.5602.

[12] Z. Zhao, P. Zheng, S. Xu, X. Wu, Object detection with deep learning: A review, IEEE Trans. Neural Netw. Learn. Syst. 30 (11) (2019) 3212–3232.

[13] L. Deng, G. Hinton, B. Kingsbury, New types of deep neural network learning for speech recognition and related applications: an overview, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 8599–8603.

[14] M.M. Lopez, J. Kalita, Deep learning applied to nlp, 2017, arXiv:1703.03091.

[15] V. Këpuska, G. Bohouta, Next-generation of virtual personal assistants (microsoft cortana, apple siri, amazon alexa and google home), in: 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), 2018, pp. 99–103.

[16] M.A. Al-Garadi, A. Mohamed, A.K. Al-Ali, X. Du, I. Ali, M. Guizani, A survey of machine and deep learning methods for internet of things (iot) security, IEEE Commun. Surv. Tutor. 22 (3) (2020) 1646–1685.

[17] X. Lu, Q. Li, Z. Qu, P. Hui, Privacy information security classification study in internet of things, in: 2014 International Conference on Identification, Information and Knowledge in the Internet of Things, 2014, pp. 162–165.

[18] K. Ren, T. Zheng, Z. Qin, X. Liu, Adversarial attacks and defenses in deep learning, Engineering 6 (3) (2020) 346–360.

[19] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, 2014, arXiv:1312.6199.

[20] P. Tabacof, E. Valle, Exploring the space of adversarial images, 2016, arXiv:1510.05328.

[21] I.J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, 2015, arXiv:1412.6572.

[22] A. Kurakin, I. Goodfellow, S. Bengio, Adversarial examples in the physical world, 2017, arXiv:1607.02533.

[23] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z.B. Celik, A. Swami, The limitations of deep learning in adversarial settings, 2015, arXiv:1511.07528.

[24] S. Moosavi-Dezfooli, A. Fawzi, P. Frossard, Deepfool: A simple and accurate method to fool deep neural networks, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2574–2582.

[25] N. Carlini, D. Wagner, Towards evaluating the robustness of neural networks, in: 2017 IEEE Symposium on Security and Privacy (SP), 2017, pp. 39–57.

[26] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, C.-J. Hsieh, Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models, in: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, AISec '17, Association for Computing Machinery, New York, NY, USA, 2017, pp. 15–26, http://dx.doi.org/10.1145/3128572.3140448.

[27] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, P. Frossard, Universal adversarial perturbations, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[28] M.A. Amanullah, R.A.A. Habeeb, F.H. Nasaruddin, A. Gani, E. Ahmed, A.S.M. Nainar, N.M. Akim, M. Imran, Deep learning and big data technologies for iot security, Comput. Commun. 151 (2020) 495–517.

[29] T. Dreossi, S. Jha, S.A. Seshia, Semantic Adversarial Deep Learning, vol. 10981, Springer International Publishing, 2018, pp. 3–26, http://link.springer.com/10.1007/978-3-319-96145-3_1.

[30] J. Su, D.V. Vargas, K. Sakurai, One pixel attack for fooling deep neural networks, IEEE Trans. Evol. Comput. 23 (5) (2019) 828–841, http://dx.doi.org/10.1109/TEVC.2019.2890858.
[31] D. Chen, R. Xu, B. Han, Patch selection denoiser: An effective approach defending against one-pixel attacks, Commun. Comput. Inf. Sci. Neural Inf. Process. (2019) 286–296.
[32] Z.-Y. Liu, P.S. Wang, S.-C. Hsiao, R. Tso, Defense against n-pixel attacks based on image reconstruction, in: Proceedings of the 8th International Workshop on Security in Blockchain and Cloud Computing, SBC '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 3–7, https://doi-org.ezproxy-f.deakin.edu.au/10.1145/3384942.3406867.
[33] S.A.A. Shah, M. Bougre, N. Akhtar, M. Bennamoun, L. Zhang, Efficient detection of pixel-level adversarial attacks, in: 2020 IEEE International Conference on Image Processing (ICIP), 2020, pp. 718–722.
[34] W. Yu, F. Liang, X. He, W.G. Hatcher, C. Lu, J. Lin, X. Yang, A survey on the edge computing for the internet of things, IEEE Access 6 (2018) 6900–6919.
[35] B. Chen, J. Wan, A. Celesti, D. Li, H. Abbas, Q. Zhang, Edge computing in iot-based manufacturing, IEEE Commun. Mag. 56 (9) (2018) 103–109.
[36] I.T. Jolliffe, J. Cadima, Principal component analysis: a review and recent developments, Phil. Trans. R. Soc. A 374 (2065) (2016) 20150202.
[37] J. Wright, Y. Peng, Y. Ma, A. Ganesh, S. Rao, Robust principal component analysis: Exact recovery of corrupted low-rank matrices by convex optimization, in: Proceedings of the 22nd International Conference on Neural Information Processing Systems, NIPS'09, Curran Associates Inc., Red Hook, NY, USA, 2009, pp. 2080–2088.
[38] H. Gfrerer, J. Ye, J. Zhou, Second-order optimality conditions for non-convex set-constrained optimization problems, 2020, arXiv:1911.04076.
[39] A.I. Chen, A. Ozdaglar, A fast distributed proximal-gradient method, in: 2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton), 2012, pp. 601–608.
[40] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, Y. Ma, Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix, 2009.
[41] https://www.cs.toronto.edu/~kriz/cifar.html. (Accessed 27 February 2021) [link].
[42] Welcome to the ini benchmark website! https://benchmark.ini.rub.de/.
[43] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324.
[44] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2015, arXiv:1512.03385.

Muhammad Akbar Husnoo is currently a PhD scholar at Deakin University. He received his dual B.Sc. (Hons) in Software Engineering from both Staffordshire University, UK and Asia Pacific University of Technology & Innovation, Malaysia in 2019. He also recently completed a Master of Data Science at Deakin University, Burwood, VIC, Australia in 2021. He has participated in several hackathons and is the 'Champion Winner' of the SAS Malaysia FinTech Competition 2017–2018. Furthermore, he has been awarded 'The University Prize for Best Project of the B.Sc. (Hons) in Software Engineering Award 2018/2019' for his honours thesis. Moreover, he was awarded the Deakin International Meritorious Scholarship for his master degree, the CSRI 2020 Summer Scholarship and a full Deakin University Postgraduate Research Scholarship to pursue his doctorate. His research interests include privacy preservation, adversarial learning, deep learning, machine learning and other related topics.

Adnan Anwar is a Lecturer and Deputy Director of postgraduate cybersecurity studies at the School of Information Technology. Previously he has worked as a Data Scientist at Flow Power. He has over 8 years of research, and teaching experience in universities and research labs including NICTA, La Trobe University, and University of New South Wales. He received his PhD and Master by Research degree from UNSW. He is broadly interested in the security research for critical infrastructures including Smart Energy Grid, SCADA system, and application of machine learning and optimization techniques to solve cyber security issues for industrial systems. He has been the recipient of several awards including UPA scholarship, UNSW TFR scholarship, best paper award and several travel grants including ACM and Postgraduate Research Student Support (PRSS) travel grants. He has authored over 40 articles including high-impact journals (mostly in Q1), conference articles and book chapters in prestigious venues. He is an active member of IEEE for over 9 years and serving different committees.