

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/328758478>

QuSecNets: Quantization-based Defense Mechanism for Securing Deep Neural Network against Adversarial Attacks

Preprint · November 2018

CITATIONS

0

READS

15

7 authors, including:



Faiq Khalid

TU Wien

44 PUBLICATIONS 79 CITATIONS

SEE PROFILE



Muhammad Shafique

Karlsruhe Institute of Technology

261 PUBLICATIONS 2,835 CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:



Towards Precise, Scalable and Automatic Analysis of Analog and Mixed Signal Circuits [View project](#)



RISPP: A run-time adaptive reconfigurable processor [View project](#)

QuSecNets: Quantization-based Defense Mechanism for Securing Deep Neural Network against Adversarial Attacks

Hassan Ali*, Hammad Tariq*, Muhammad Abdullah Hanif†, Faiq Khalid†, Semeen Rehman†, Rehan Ahmed* and Muhammad Shafique†

* National University of Science and Technology, Islamabad, Pakistan

† Vienna University of Technology (TU Wien), Vienna, Austria

{hali.msee17seecs, htariq.msee17seecs, rehan.ahmed}@seecs.edu.pk, {muhammad.hanif, faiq.khalid, seemeen.rehman, muhammad.shafique}@tuwien.ac.at,

Abstract — Deep Neural Networks (DNNs) have recently been shown vulnerable to adversarial attacks in which the input examples are perturbed to fool these DNNs towards confidence reduction and (targeted or random) misclassification. In this paper, we demonstrate that how an efficient quantization technique can be leveraged to increase the robustness of a given DNN against adversarial attacks. We present two quantization-based defense mechanisms, namely Constant Quantization (CQ) and Variable Quantization (VQ), applied at the input to increase the robustness of DNNs. In CQ, the intensity of the input pixel is quantized according to the number of quantization levels. While in VQ, the quantization levels are recursively updated during the training phase, thereby providing a stronger defense mechanism. We apply our techniques on the Convolutional Neural Networks (CNNs, a particular type of DNN which is heavily used in vision-based applications) against adversarial attacks from the open-source *Cleverhans* library. Our experimental results show 1%-5% increase in the adversarial accuracy for MNIST and 0%-2.4% increase in the adversarial accuracy for CIFAR10.

Keywords— Machine Learning, DNN, Quantization, Variable Quantization, Security, Adversarial Machine Learning, Defense.

I. INTRODUCTION

Over the past few years, machine learning, especially, deep neural networks (DNNs)-based computing paradigms have been emerged as prime solution for handling the enormous amount of data in many applications, i.e., autonomous vehicles, healthcare, transportation management, etc. [1]. Though, DNNs demonstrate the tremendous success in addressing several computing challenges, but these algorithms are vulnerable to several security threats [18]. The reason behind these security vulnerabilities in DNNs is their dependencies on input datasets, e.g., training of these models is heavily dependent on the training dataset. This data dependency can be exploited to perform several security attacks, i.e., data poisoning during ML training or inference. However, data poisoning during the training process requires complete access to data, training process and baseline ML algorithm. Moreover, these data poisoning attacks have a direct impact on a trained ML algorithm which can be exploited to neutralize these attacks.

On the other hand, data poisoning attacks during the inference, usually do not change the trained ML algorithm but it exploits their data dependencies to perform security attacks, i.e., misclassification and confidence reduction. However, the key challenge in manipulating the inference data is imperceptibility of added noise. For example, adversarial examples generate the imperceptible adversarial noise to fool the DNN inferencing [3]. Therefore, there is a dire need of developing the defense mechanisms to protect the DNN inference.

Several defense mechanisms against adversarial attacks have been proposed, i.e., adversarial training [10][13][12] and DNN masking [14][15]. Though, these defense mechanisms either change the DNN structure or train it against the known adversarial vulnerabilities, which limits their defense scope to known vulnerabilities. Moreover, several counter-attacks have proposed to neutralize these defense mechanisms [6][16][17]. Therefore, more flexible and comprehensive defense strategy is required that can reduce the imperceptibility to make it detectable.

A. Motivational Analysis

As an alternative solution, quantization has emerged as one of the prime solutions to reduce the imperceptibility of adversarial examples [12][15] because of their inherent property of being insensitive to small perturbations. To show this behavior, we perform an experiment on one of the commonly used CNN structure and adversarial attacks from the open-source *Cleverhans* [19] library by providing it the quantized clear and perturbed images. Fig. 1 shows that just by introducing the quantization in an attack image its adversarial accuracy significantly increased to (24.66% from 8.87%). It can also be observed that the quantized adversarial images are almost identical to the quantized clean images.





Clean						
						
Image	Real	Q2	Q3	Real	Q2	Q3
Accuracy	8.87	24.66	19.6	8.87	24.66	19.6

Fig. 1: Comparison of the quantized output for clean and adversarial (Adv.) images of CIFAR10

Table 1: Summary of the Existing Defense Strategies for Machine Learning Networks

Techniques	Approaches	Related Work	Threat Model	Description	Motivation	Tunable Defense Parameters	Limitations
<i>Feature Squeezing</i>	Input Transformation	[12]	White-box	Reduces the number of bits in order to make the inputs insensitive to small perturbations	Improved Accuracy	No	JSMA, CW
			Black-box				
<i>BReLU + GDA</i>	Input Transformation	[14]	White-box	Instead of ReLU, introduces and uses Bounded ReLU activation and Augments the input data with the Gaussian Noise during training	Improved Accuracy	No	JSMA, CW
	Adversarial Learning		Black-box				
<i>APE-GAN</i>	Input Transformation	[13]	White-box	Introduces a GAN as the input to the CNN. GAN is trained to generate clean images out of both clean and Adversarial images	GAN based defense	Yes	Attacks on GAN
	Adversarial Learning		Black-box				
<i>Defense-GAN</i>	Input Transformation	[20]	White-box	Uses a trained GAN with random inputs to generate an image and minimizes the mean square error of input and generated image	GAN based defense	Yes	Attacks on GAN
			Black-box				
<i>Adversarial Training</i>	Adversarial Learning	[10]	White-box	Trains the DNN on adversarial examples crafted out of Training dataset	Robustness is embedded within the DNN	No	Unknown Attacks
			Black-box				
<i>Defensive Distillation</i>	Gradient Masking	[11]	White-box	Trains the DNN based on the outputs of an already trained DNN	Robustness is embedded with the DNN	No	Transfer Attacks
					Smooth Label training		
<i>Dynamic Quantization Activation</i>	Gradient Masking	[15]	White-box	Train the DNN with a Dynamic Quantization Layer at the input of the Conventional DNN and introduces a tunable threshold	Quantization with tunable Thresholds	Yes	Transfer Attacks
	Input Transformation		Black-box				Black-box Attacks
							JSMA, CW
<i>QuSecNets</i>	Input Transformation	This paper	White-box	Trains the DNN with a differentiable variable threshold Quantization Layer and constant threshold quantization Layer	Differentiable Quantization with tunable thresholds	Yes	JSMA, CW
	Adversarial Learning		Black-box		Integer number of quantization levels		
					Increased Perceptibility of Adversarial Examples		

B. Associated Research Challenges

Though quantization is insensitive to small perturbations, to reduce the imperceptibility following research challenges should be addressed:

1. How to explore the quantization effects on the gradient to develop the defense strategies?
2. How to explore the quantization to reduce the imperceptibility of the adversarial examples?
3. How to address the flexibility problem with respect to quantization levels in both variable and constant quantization techniques?

C. Novel Contribution

To address the above-mentioned research challenges, we propose to *leverage constant and variable quantization to develop the defense against adversarial attacks*. In summary, in this paper, we make the following contributions:

1) **QuSecNets (Section III):** We propose to leverage insensitive nature of quantization towards small perturbations by *introducing an additional trainable quantization layer at the input of a DNN*.

2) **Trainable Quantization (Section III.A):** We propose to exploit the tunable thresholds in variable quantization to train the quantization layer against several types of adversarial attacks.

II. STATE-OF-THE-ART DEFENSES AGAINST ADVERSARIAL ATTACKS

In this section, we provide the brief explanation of the adversarial examples and the state-of-the-art defense mechanisms. Table 1 provides a comprehensive view of different related works in a categorized fashion highlighting their key features, threat models, and limitations. In the following, we discuss the most relevant works in more detail.

A. Adversarial Examples

Adversarial examples are usually small and invisible perturbation in the input from adversary to manipulate the ML-based classification to perform the targeted or un-targeted attacks on a DNN. The imperceptibility and strength of these attacks are highly dependent on their optimization function. Therefore, based upon the attack strategies, these approaches can be categorized as follows [1].

Gradient-Based Attacks: These attacks generate the adversarial noise based on the gradient of the loss function with respect to the corresponding parts of the input images/samples. Some of the most commonly used gradient-based attacks are Fast Gradient Sign Method (FGSM) [3], Jacobian Saliency Map Attack (JSMA) [4], Basic Iterative Method (BIM) [5], Carlini-Wagner Attacks (CW) [6] and DeepFool [7].

Decision Based Attacks: Unlike Gradient-based attacks, Decision Based attacks do not require the calculation or estimation of gradients of loss. Instead, they utilize the class decision of the model in order to introduce the perturbation in the input image. Examples include Point-wise Attack [8] and Additive Gaussian Noise Attack [8].

Score Based Attacks: These attacks analyze the statistical or probabilistic behavior of the individual input components to estimate the corresponding gradients with respect to loss function. Some of the common used attacks are Single-Pixel Attack [9] and Local Search Attack [9].

B. Defense Mechanisms

Adversarial attacks are one of the major security vulnerabilities in ML-based applications. Therefore, to improve and ensure the security of ML-based applications several defense strategies have been proposed based on the DNN masking [12], gradient masking [11], training for known adversarial attacks [10][13] and quantization of input of the DNNs [12][15].

One of the most commonly used approach is to train the DNNs for the adversarial attacks. For example, *adversarial learning* trains the DNN based on the known adversarial examples but it limits its scope to known adversarial attack. Based on the similar concept, Adversarial Perturbation Elimination using GANs (APE-GAN) has been proposed. This technique considers each input sample as a potential adversarial example and retrain the network to remove the imperceptible noises. However, by introducing relatively stronger adversarial noise than the one at which APE-GAN is trained, one can break this defense mechanism.

Another approach is to either *mask the gradient or the whole DNN*, e.g., Defensive Distillation masks the gradient of the network, but it is only valid for gradient-based attacks and it can be compromised by empirically inferring the gradient by applying different loss functions [16]. To address this issue, several techniques have been proposed to mask the entire DNN to limit the leakage of empirical information. The activation function can also be exploited to avoid small perturbations, i.e., Bounded Relu and Gaussian Data Augmentation (BReLU+GDA) [14].

As an alternative solution, *quantization* has emerged as one of the prime defense mechanisms against adversarial examples. For example, feature squeezing technique uses the binary quantization [12]. Similarly, the dynamic quantization techniques have been explored to neutralize the impact of adversarial example [15]. *However, due to less flexibility in quantization levels for both constant and variable quantization, these techniques are limited to known attacks. Therefore, in this paper, we leverage integral quantization level and flexibility variable quantization to develop a flexible and trainable defense mechanism.*

III. QUANTIZATION-BASED DEFENSE FOR DNN

In this section, we discuss the proposed methodology, QuSecNets, which exploits the quantization to develop a defense mechanism against the adversarial attack in DNNs. The proposed methodology consists of the following steps, as shown in Fig. 2:

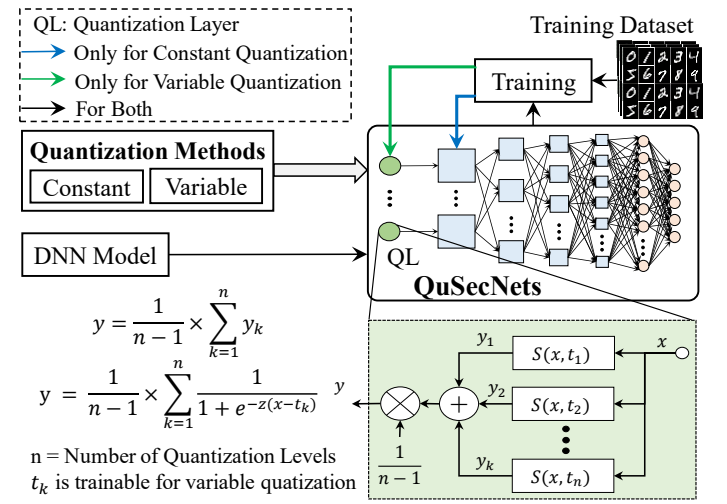


Fig. 2: Proposed Methodology to design quantization-aware secure DNNs

1. First, we select a DNN model and the quantization methods based on the targeted vulnerabilities, input dataset and applications.
2. Afterwards, we design and integrate the quantization layer within the selected DNN model. This quantization layer has one-to-one relation with the input layer, i.e., for each input pixel there is one quantization function. This function is defined as the average value of “n” sigmoid function “ $S(x, t_k)$ ” (See Fig. 2) and can be formulated as:

$$y = S(x, t_k) = \frac{1}{n-1} \times \sum_{k=1}^n \frac{1}{1 + e^{-z(x-t_k)}} \quad (1)$$

Where x , z and t_k is the intensity of a single pixel from input image, the scalar constant and the threshold values which *may* (in variable quantization) or *may not* (in constant quantization) be trained, respectively. Moreover, the n represents the number of quantization levels, e.g., if the number of quantization levels is 2, all the values below 0.5 are set to 0, while those above or equal to the 0.5 are set to 1

3. After integrating it with the DNN, we train the modified network, depending upon the quantization methodology. For example, in case of constant quantization, we do not train the quantization layer. However, in case of variable quantization we propose to utilize the backpropagation methodology.

A. Training the Quantization Layer

In case of variable quantization, we need to train the thresholds (t_k) based on the output prediction/classification probabilities of the DNNs. Therefore, first we define a cost function to identify the difference between the actual and targeted predictions.

$$cost = \frac{1}{c} \times \sum_{c=1}^c (P(x=c) - P(x=c_{acc}))^2 \quad (2)$$

Where, $P(x=c_{acc})$ and $P(x=c)$ represent the probability of input x to be classified as any class c and actual class c_{acc} , respectively. Similar to the back-propagation algorithm in DNN training, the threshold of the quantization layer can be trained by minimizing the cost function (Equation (3)):

$$t_{k,new} = t_{k,old} - \eta \frac{\partial (cost)}{\partial t_{k,old}} \quad (3)$$

Where

$$\frac{\partial (cost)}{\partial t_{k,old}} = \frac{\partial (cost)}{\partial y} \times \frac{\partial (y)}{\partial t_{k,old}} \quad (4)$$

$$\frac{\partial (y)}{\partial t_{k,old}} = \left(\sum_{all\ k\ in\ y} \left(\frac{-1}{n-1} \right) (z)(y_k)(1-y_k) \right) \quad (5)$$

$$\frac{\partial (cost)}{\partial y_k} = \sum_{all\ n} w_{nk} \delta_n = \Delta_k \quad (6)$$

Where, y_k is the output of the sigmoid function and " δ_k " denotes the sensitivity of y_k for all the variables in the first convolutional layer of the DNN defined by the backpropagation. Thus, the change in cost function is determined by the following equation.

$$\frac{\partial (cost)}{\partial t_{k,old}} = \left(\sum_{all\ k\ in\ y} \Delta_k \right) \times \left(\sum_{all\ k\ in\ y} \left(\frac{-1}{n-1} \right) (z)(y_k)(1-y_k) \right) \quad (7)$$

By combining the Equations 3 and 7, the training equation for quantization threshold modeled as:

$$t_{k,new} = t_{k,old} - \eta \times \left(\sum_{all\ k\ in\ y} \Delta_k \right) \times \left(\sum_{all\ k\ in\ y} \left(\frac{-1}{n-1} \right) (z)(y_k)(1-y_k) \right) \quad (8)$$

IV. EXPERIMENTAL RESULTS

To illustrate the effectiveness of our *QuSecNets*, we integrate the quantization layer with one of the commonly used baseline CNN structure from the open-source *Cleverhans* library which also has several online adversarial attacks. We perform several analyses using the experimental setup shown in Fig. 4.

- **CNN:** we use the following CNN structure from the *Cleverhans* library: *Conv2D(64, 8x8) - Conv2D(128, 6x6) - Conv2D(128, 5x5) - Dense(10) - Softmax()*.
- **Dataset:** we trained and tested the above mention CNN structure for the CIFAR10 and the MNIST datasets with experimental parameters given in Table 2.

Table 2: Experimental parameters for CNN training and testing

Parameters	Value
No. of epochs	6 -10
Number of Images in Test Dataset	10000
Batch Size	128
Learning Rate	0.001

- **Adversarial Attack:** we implemented some of the state-of-the-art adversarial attacks, i.e., FGSM, JSMA, and CW-L2.
- **Threat Model:** We assume both White-box and Black-box threat models. In both models, the adversary can only alter the inputs, however, in white box scenario the adversary knows and uses the weights and biases of our *QuSecNet*.

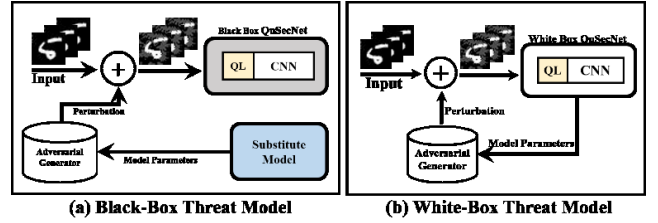


Fig. 3: Attacking the QuSecNet assuming a) Black-box Threat Model and b) White-box Threat Model.

A. Experimental Analysis

This section provides the detailed analysis of the *QuSecNets* against the implemented adversarial attacks.

Impact of Variable Quantization on Adversarial Training: It can be observed that Variable Quantization Layer, though quite ineffective as a solo defense highly assists adversarial training of a network (see Fig. 5). Authors of [15] also made this observation but their evaluation seems to have a major flaw. The impact of Dynamic Quantization Activation (DQA) on Adversarial Training, using gradient-based attacks, gives delusional results as the "signum" function masks the gradient and the adversary is unable to generate adversarial examples (see Fig. 6). Considering this fact, the positive impact of DQA on Adversarial Training becomes unclear.

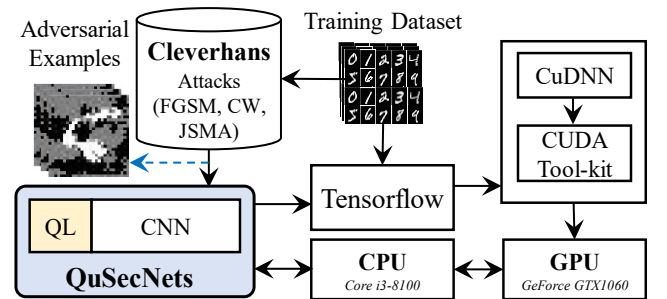


Fig. 4: Experimental analysis of QuSecNets against multiple attacks from Cleverhans for MNIST and CIFAR10 dataset.

Effect of "z" - the Scalar Constant: In this experimental analysis, the value of scalar constant "z" is set to be 5 for Variable Quantization Layer. This parameter is directly proportional to the

quantization effect. However, for higher values it reduces derivative slope of quantized output and makes it a constant function which halts the training process. This effect is shown in Equation (9).

$$t_{1,new} = t_{1,old} - \eta \times \frac{d}{d y} cost \times \frac{z \times -y(1-y)}{n-1} \quad (9)$$

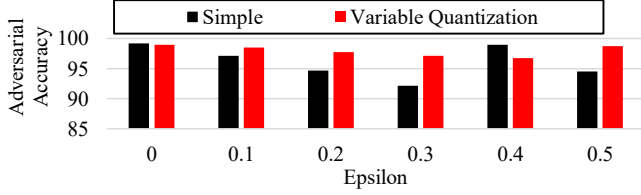


Fig. 5: Accuracy on adversarial examples generated by FGSM for Adversarial Training of simple CNN vs. Adversarial Training of Variable QuSecNet for MNIST

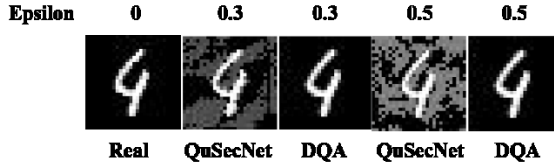


Fig. 6: Comparison of Adversarial Examples generated for QuSecNets and DQA [14] with clean image.

Through experimental analysis, we observed that value of “z” should be set between 5 and 40 to train the variable quantization layer effectively. However, for the constant quantization, a higher value (we use $z = 50$) is required to achieve better quantization effects.

Effect of changing the number of Quantization Levels in Quantization Layers:

Increasing the quantization levels tends to decrease the output accuracy as shown in the Fig. 7. This is justifiable as increasing the quantization levels effectively makes the quantized input image more similar to the actual non-quantized input image. In addition, an increase in quantization levels also increases the number of transitions in the Quantization layer. This makes the input more sensitive to the adversarial noise.

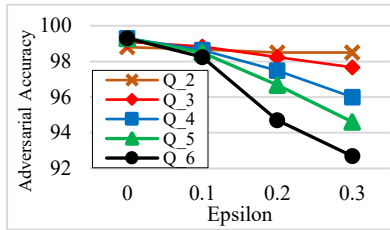


Fig. 7: Adversarial Accuracy of Constant QuSecNets against FGSM attack for different values of Epsilon

Increased Perceptibility of Adversarial Examples: Constant Quantization and Variable Quantization tend to reduce the imperceptibility of the Adversarial Examples (See Fig. 8). This is due to reduced insensitivity of the QuSecNet to small changes in the inputs.

We observe that the quantization showed better results in terms of Adversarial Accuracy for MNIST as compared to CIFAR10

(See Fig. 9 to Fig. 11). We believe that this is due to the clustered distribution of MNIST.



Fig. 8: Adversarial examples generated to fool QuSecNets for various defense strategies. Clean images are given for comparison.

V. COMPARISON WITH STATE-OF-THE-ART DEFENSE MECHANISMS

To demonstrate the effectiveness of the proposed defense mechanism, in this section, we present a comparison with state-of-the-art defense mechanisms, i.e., Feature Squeezing and BReLU+GDA.

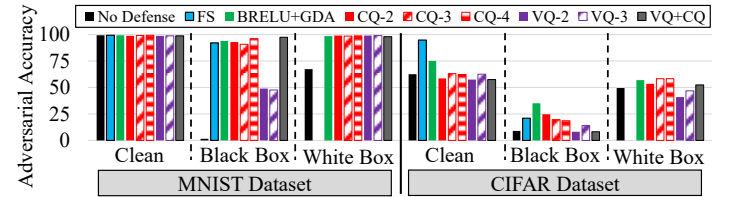


Fig. 9: Comparison of the proposed strategies with the state-of-the-art defenses for FGSM attacks. Epsilon = 0.3 for MNIST and Epsilon = 0.1 for CIFAR10

Fig. 9 reports our results for the FGSM attacks against different Defense mechanisms on MNIST and CIFAR10. The adversarial accuracy captures the accuracy of the DNNs on adversarial examples. Clean accuracies for the MNIST are observed to be reduced by a very small amount for Constant and Variable Quantization. The constant quantization works most effectively for the white-box FGSM attacks (from 1.48% to 97.51%). BReLU [14] reports an accuracy of 94% for such attacks) due to the insensitivity of the Quantization Layer to small perturbations in inputs. However, these results are not as appealing for CIFAR10 white box FGSM attacks. We believe this is because of the clustered distribution of the MNIST dataset, which increases the effectiveness of the Quantization Layer. We observe that integral constant quantization almost always outperforms every other defense strategy including the state-of-the-art defenses (Feature Squeezing [12] and BReLU [14]). Variable Quantization surpasses the Constant Quantization if the Adversarial Training is performed for the DNN (See Fig. 5 and Fig. 9). Our white-box results against FGSM attacks for CIFAR10 are inferior to BReLU. However, for the black-box attacks our defenses compete and often surpass the BReLU in effectiveness.

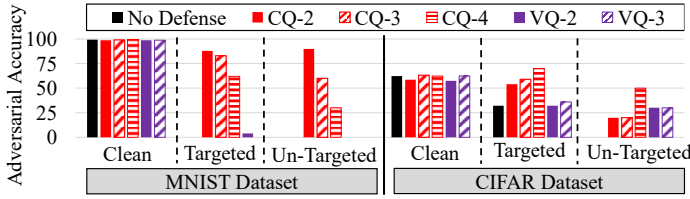


Fig. 10: Comparison of the proposed strategies for CW-L2 attacks. No. of source samples for targeted CW-L2 attacks = 10, No. of iterations = 100.

Results for targeted and untargeted CW-L2 attacks for different defense strategies are reported in Fig. 10 for comparison. QuSecNets cause significant improvement in the robustness of DNNs for both MNIST (88% for white-box and 90% for Black-box) and CIFAR (54% for White-box and 50% for CIFAR) as compared to CNNs with no defense mechanism (0%). We do not report the results for Feature Squeezing and BReLU here, because of different DNN Models used for evaluation of results.

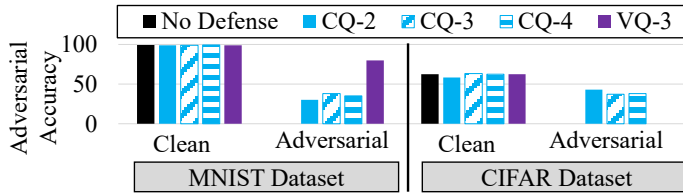


Fig. 11: Comparison of the proposed for Targeted JSMA attacks. No. of source samples = 10, No. of iterations = 100.

Fig. 11 reports the results of QuSecNets against targeted JSMA attacks. The reason for the inferior results is that JSMA introduces concentrated noise at few pixels in the input image instead of small distributed perturbation, which the Quantization Layer effectively counters. The accuracy for concentrated attacks can be significantly improved by other strategies such as median filters [12] or input drop-out.

VI. CONCLUSION

In this paper, we propose to leverage the insensitive and dynamic nature of constant/variable quantization towards small perturbation for developing a defense mechanism, QuSecNets. This methodology introduces an additional layer at the input of DNNs, to reduce the imperceptibility. To demonstrate the effectiveness of the proposed methodology, we evaluate our approach against some of the state-of-the-art attacks, i.e., FGSM, CW, JSMA and compare it with the state-of-the-art. We empirically prove that Integral Constant Quantization Layer significantly hardens the DNNs. In addition, we show that QuSecNets have a better Adversarial Learning capability as compared to the conventional DNNs for clustered datasets.

ACKNOWLEDGEMENT

This work is partly supported by the Erasmus+ Motility Program between Vienna University of Technology (TU Wien), Austria and National University and Technology Pakistan.

REFERENCES

- [1] Stilgoe, J., 2018. Machine learning, social learning and the governance of self-driving cars. *Social studies of science*, 48(1), pp.25-56.
- [2] J. Rauber, W. Brendel, and M. Bethge, "Foolbox v0.8.0: A python toolbox to benchmark the robustness of machine learning models," arXiv preprint arXiv:1707.04131, 2017. [Online]. Available: <http://arxiv.org/abs/1707.04131>
- [3] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and Harnessing Adversarial Examples." In International Conference on Learning Representations (ICLR), 2015.
- [4] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. "The Limitations of Deep Learning in Adversarial Settings." In IEEE European S&P, 2016.
- [5] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. "Adversarial examples in the physical world." CoRR, abs/1607.02533, 2016. URL <http://arxiv.org/abs/1607.02533>.
- [6] Nicholas Carlini and David Wagner. "Towards evaluating the robustness of neural networks." In IEEE Symposium on Security and Privacy, 2017.
- [7] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. "DeepFool: a simple and accurate method to fool deep neural networks." In IEEE CVPR, 2016.
- [8] J. Rauber, W. Brendel, and M. Bethge, "Foolbox v0.8.0: A python toolbox to benchmark the robustness of machine learning models," arXiv preprint arXiv:1707.04131, 2017. [Online]. Available: <http://arxiv.org/abs/1707.04131>
- [9] Naroditska, Nina and Kasiviswanathan, Shiva Prasad. "Simple black box adversarial perturbations for deep neural networks." CoRR, abs/1612.06299, 2016. URL <http://arxiv.org/abs/1612.06299>.
- [10] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. "Intriguing properties of neural networks." In International Conference on Learning Representations (ICLR), 2014.
- [11] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. "Distillation as a defense to adversarial perturbations against deep neural networks." In IEEE Symposium on Security and Privacy, 2016.
- [12] Weilin Xu, David Evans, and Yanjun Qi. "Feature squeezing: Detecting adversarial examples in deep neural networks." CoRR, abs/1704.01155, 2017. URL <http://arxiv.org/abs/1704.01155>
- [13] S. Shen, G. Jin, K. Gao, and Y. Zhang. "APE-GAN: Adversarial Perturbation Elimination with GAN." arXiv:1707.05474, 2017.
- [14] V. Zantedeschi, M.-I. Nicolae, and A. Rawat. "Efficient defenses against adversarial attacks." arXiv preprint arXiv:1707.06728, 2017.
- [15] Adnan Siraj Rakin, Jinfeng Yi, Boqing Gong, Delian Fang, "Defend Deep Neural Networks Against Adversarial Examples via Fixed and Dynamic Quantized Activation Functions", URL <https://arxiv.org/pdf/1807.06714>.
- [16] Nicholas Carlini and David Wagner. "Defensive distillation is not robust to adversarial examples." CoRR, abs/1607.04311, 2016. URL <http://arxiv.org/abs/1607.04311>
- [17] Nicholas Carlini and David Wagner, "Magnet and "Efficient Defenses against Adversarial Attacks" are not Robust to Adversarial Examples", URL <https://arxiv.org/abs/1711.08478>
- [18] Shafique, M., Theodoridis, T., Bouganis, C.S., Hanif, M.A., Khalid, F., Hafiz, R. and Rehman, S., 2018, March. "An overview of next-generation architectures for machine learning: Roadmap, opportunities and challenges in the IoT era". In Design, Automation & Test in Europe Conference & Exhibition (DATE), 2018 (pp. 827-832). IEEE.
- [19] Papernot, Nicolas, et al. "cleverhans v2. 0.0: an adversarial machine learning library." arXiv preprint arXiv:1610.00768 (2016).
- [20] Pouya Samangouei, Maya Kabkab, and Rama Chellappa, "Defense-GAN: Protecting Classifiers against Adversarial Attacks using Generative Models" URL <https://arxiv.org/pdf/1805.06605>