EFFICIENT RANDOMIZED DEFENSE AGAINST ADVERSARIAL ATTACKS IN DEEP CONVOLUTIONAL NEURAL NETWORKS

Fatemeh Sheikholeslami[1], Swayambhoo Jain[2], and Georgios B. Giannakis[1]

[1]University of Minnesota - Twin Cities, USA, [2]Technicolor AI Lab - Palo Alto, USA

ABSTRACT

Despite their well-documented learning capabilities in clean environments, deep convolutional neural networks (CNNs) are extremely fragile in adversarial settings, where carefully crafted perturbations created by an attacker can easily disrupt the task at hand. Numerous methods have been proposed for designing effective attacks, while the design of effective defense schemes is still an open area. This work leverages randomization-based defense schemes to introduce a sampling mechanism for strong and efficient defense. To this end, sampling is proposed to take place over the matricized mid-layer data in the neural network, and the sampling probabilities are systematically obtained via variance minimization. The proposed defense only requires adding sampling blocks to the network in the inference phase without extra overhead in the training. In addition, it can be utilized on any pre-trained network without altering the weights. Numerical tests corroborate the improved defense against various attack schemes in comparison with state-of-the-art randomized defenses.

Index Terms— Deep learning, convolutional neural networks, adversarial examples, randomized defenses, image classification.

1. INTRODUCTION

Deep Neural Networks (DNNs) gained increasing popularity as their capability in diverse tasks such as object recognition and detection [1, 2], speech recognition and language translation [3], voice synthesis [4], and many more, reach or even surpass human-level accuracy. However, recent studies have cast doubt on the reliability of DNNs as highly-accurate networks are shown to be extremely vulnerable to carefully crafted inputs designed to fool them [5, 6]. This will challenge applicability of the DNNs in terms of safety and security in critical environments such as autonomous cars [7], automatic speech recognition [8], and face detection [9, 10].

Particularly, in the case of convolutional neural networks (CNN) for image classification, the severity of brittleness is highlighted because small adversarial perturbations on the clean data are often imperceptible to the human eye, however they can cause the trained CNNs to classify the *adversarial examples* incorrectly with high confidence. Furthermore, adversarial noise generated using a given trained network can successfully fool another CNN-based classifier [11]. This addresses practical *black-box* attacks, where the attacker does not have access to the target classifier, however he/she has a high chance of sabotage. Thus, improving the robustness of CNNs is of high importance for real-world applications in potentially adversarial settings, especially in sensitive applications.

Majority of this work was done during a summer research internship at Technicolor AI Lab in Palo Alto, CA - USA. This research was supported in part by NSF grant 1500713, 151405\6, 1505970, and 1711471. Author emails: sheik081@umn.edu, swayambhoo.jain@technicolor.com, georgios@umn.edu

Design of powerful adversarial perturbations in environments with different levels of knowledge about the target CNN as well as affordable complexity, have been considered in numerous works [6, 12, 13]. Similarly, design of defense methods for enhancing the robustness of CNNs against adversarial perturbations has pursued two broad directions of *detection* and attack *recovery* schemes. The defense mechanism for the former aims at detection of adversarial images by classifying the input images into clean or adversarial ones, by utilizing different tools such as auto-encoders [14], detection subnetwork learned during the training phase [15, 16], and dropout units [17]. On the other hand, recovery schemes aim at enhancing the robustness of the classification accuracy by data pre-processing [18], adversarial training [19, 20], and Lipschitz regularization [21, 22] among other schemes.

Along the objective of this work and by focusing on attack detection schemes, it has been shown that randomization-based defenses exhibit higher robustness against strong attacks, while other defense mechanisms can easily fail [12]. In particular, dropout units have been analyzed from a Bayesian point of view in [23], where it has been shown that they can provide a measure of *(un)certainty* on the classification output. Subsequently, [17] utilizes randomness of dropout units during the test phase as a defense mechanism, where images with high classification uncertainty are declared as adversary. Recently, randomized defense has been generalized to non-uniform sampling known [24], where mid-layer tensors are vectorized and randomly sampled, with probabilities proportional to the entry values.

Inspired by [17, 24], the goal here is to provide a systematic approach for obtaining an optimal and more efficient sampling scheme, where instead of vectorizing the mid-layer tensors as in [24], blocks of entries are sampled via reshaping the tensors into matrices, a.k.a. *matricization*, for a faster inference. This is motivated by leveraging the structure of the tensor image, where the sampling is in fact selecting fibers of the 3D tensor, corresponding to *pixels* across different *filters* in the CNN mid-layers. The sampling probabilities are then obtained by casting the problem as a variance minimization, whose convexity yields an efficient solver. Numerical results corroborate the effectiveness of the proposed defense, while improving sampling efficiency due to reduced complexity of block sampling. The advantages of our novel method can be summarized as follows.

- The proposed sampling unit can be placed in any network regardless of its size and depth.
- The defense scheme takes place in the test phase, thus imposing no overhead in the training phase while also keeping trained weights of the network untouched.
- Educated and structured sampling is utilized, where blocks of data are sampled via optimally learned probabilities, thus increasing sampling efficiency as well as defense strength.

2. MATRICIZED VARIANCE MINIMIZATION DEFENSE

In this work, we build on the randomized defense schemes, as the introduced randomness enables measuring the (un)certainty of the output class as a means to detecting adversarial images. However, unlike simple dropout methods which utilize blind Bernoulli random variables, we aim at properly learning sampling probabilities to improve performance. Furthermore, as also addressed in [23], the dropout (sampling) unit may be placed at any point in the CNN architecture, including the input image itself. Thus, to enhance interpretability, we will build our objective by focusing on sampling the input image first. The objective will be readily generalized later for the hidden layers of the network.

Most adversarial perturbations are additive carefully crafted noise, yielding adversarial images as

$$\underline{\mathbf{X}}_{\text{adv}} = \underline{\mathbf{X}}_{\text{clean}} + \underline{\mathbf{N}}$$

where $\underline{\mathbf{X}}_{\text{clean}}$ is the clean (tensor) image, and $\underline{\mathbf{N}}$ is the adversarial noise, both of size $m \times n \times h$. In order to utilize the inherent structure of an image for a more efficient and smarter sampling, rather than independent sampling across entries, our idea is to matricize the tensor into a matrix of size $mn \times h$, and systematically learn the row sampling probabilities $\mathbf{p} := [p_1, ..., p_{mn}]^\top$.

Upon matricization, the adversarial image can be expressed as $\mathbf{X}_{\text{adv}} = \mathbf{X}_{\text{clean}} + \mathbf{N}$, where the $m \times n \times h$ tensors are substituted by their $mn \times h$ matricized counterparts. The proposed row sampling method for sampling $\mathbf{X}_{\text{adv}}$ proceeds as follows.

For a given sampling probability vector $\mathbf{p}$, and a total number of c draws, select index $i \sim \text{categorical}(\mathbf{p})$ for c independent draws with replacement, and gather all the drawn indices in the index set $\mathcal{I}$. Note that since draws are with replacement, we have $|\mathcal{I}| \leq c$. The randomized approximation of the image is then given by $\widehat{\mathbf{X}} = \mathbf{S}\mathbf{D}\mathbf{X}_{\text{adv}}$, where $\mathbf{S}_{mn\times mn}$ is the sampling matrix with $S_{ii} = 1$ for $i \in \mathcal{I}$, and zero otherwise. Similarly, the diagonal matrix $\mathbf{D}_{mn\times mn}$ scales the selected rows $i \in \mathcal{I}$ by the factor D_{ii}.

One would ideally seek a sampling scheme such that $\mathbb{E}[\widehat{\mathbf{X}}] = \mathbf{X}_{\text{clean}}$. In lieu of such a scheme, we choose the scaling matrix $\mathbf{D}$ such that an unbiased approximation is provided for the clean image, that is, if $\mathbf{N} = \mathbf{0}$, then $\mathbb{E}[\widehat{\mathbf{X}}] = \mathbb{E}[\mathbf{S}\mathbf{D}\mathbf{X}_{\text{clean}}] = \mathbf{X}_{\text{clean}}$. To this end, scaling is selected as $D_{ii} = 1/(1-(1-p_i)^c)$, and the off-diagonal entries are set to 0. The algorithm is tabulated in Alg. 1.

This choice of matrix $\mathbf{D}$ gives rise to unbiased approximations for the noise component as well as the adversarial image, i.e. $\mathbb{E}[\mathbf{S}\mathbf{D}\mathbf{N}] = \mathbf{N}$ and $\mathbb{E}[\widehat{\mathbf{X}}] = \mathbf{X}_{\text{adv}}$. Since unbiased approximation of the clean image is unavailable, one is motivated to find the sampling probabilities $\mathbf{p}$ such that the variance of the clean image approximation is minimized; that is,

$$\min_{\mathbf{p}\geq 0, \mathbf{1}^\top\mathbf{p}=1} \mathbb{E}\left[\|\widehat{\mathbf{X}}_{\text{clean}} - \mathbf{X}_{\text{clean}}\|_F^2\right] \tag{1}$$

where $\widehat{\mathbf{X}}_{\text{clean}} = \mathbf{S}\mathbf{D}\mathbf{X}_{\text{clean}}$. Intuitively, minimizing (1) is of interest since low values of the variance of $\widehat{\mathbf{X}}_{\text{clean}}$ will make different realizations of $\widehat{\mathbf{X}}_{\text{clean}}$ to concentrate around its mean $\mathbf{X}_{\text{clean}}$ with high probability, while the same will not happen for the adversarial noise component $\mathbf{N}$.

Expanding the objective across rows, one readily obtains

$$\begin{aligned}
\mathbb{E}[\|\widehat{\mathbf{X}}_{\text{clean}} - \mathbf{X}_{\text{clean}}\|_F^2] &= \sum_{i=1}^{mn} \mathbb{E}[\|S_{ii}D_{ii}\mathbf{x}_i^{\text{clean}} - \mathbf{x}_i^{\text{clean}}\|_2^2] \\
&= \sum_{i=1}^{mn} \mathbb{E}[\|S_{ii}D_{ii}\mathbf{x}_i^{\text{clean}}\|_2^2] - \mathbb{E}[\|\mathbf{x}_i^{\text{clean}}\|_F^2] \\
&= \sum_{i=1}^{mn} \|\mathbf{x}_i^{\text{clean}}\|_2^2 \left(\frac{1}{\pi_i} - 1\right) \qquad (2)
\end{aligned}$$

where $\mathbf{X}_{\text{clean}} := [\mathbf{x}_1^{\text{clean}}, ..., \mathbf{x}_{mn}^{\text{clean}}]^\top$, and the second equality is written using the fact that draws are iid and with replacement, rendering binary random variables $S_{ii} \sim \text{Bernoulli}(\pi_i)$ where $\pi_i = 1-(1-p_i)^c$. Thus, (1) can be rewritten as

$$\min_{\mathbf{p}\geq 0, \mathbf{1}^\top\mathbf{p}=1} \sum_{i=1}^{n} \frac{1}{1-(1-p_i)^c} \|\mathbf{x}_i^{\text{clean}}\|_2^2 \,. \tag{3}$$

We refer to the probabilities obtained by solving (3) as Matricized Variance Minimization (MVM) probabilities. It is easy to show that within the feasible set of simplex vectors, the objective of (3) is convex, which together with the convexity of the feasible set render the minimization convex. The optimal value for sampling probabilities $\mathbf{p}$ can thus be obtained by a projected gradient descent solver, with smart initialization, to prevent getting stuck at local optima due to possible cases with $\|\mathbf{x}_i^{\text{clean}}\| = 0$, as tabulated in Alg. 2. Operator $\Pi_{\text{simplex}}(.)$ denotes projection onto the simplex set, $\alpha_i = \|\mathbf{x}_i^{\text{clean}}\|_2^2$, and nnz(.) denotes the number of non-zero entries.

In practice however, one only has access to the adversarially perturbed image $\mathbf{X}_{\text{adv}}$ rather than the clean $\mathbf{X}_{\text{clean}}$. This along with the

Algorithm 1: Matrix approximation via row sampling with replacement.

Input: Matrix $\mathbf{Z}$ and probabilities $\mathbf{p} = [p_1, \cdots, p_{mn}]^\top$

Initialize $\mathbf{S}, \mathbf{D} = \mathbf{0}_{mn\times mn}$, and set diagonal entries of $\mathbf{D}$ as

$$[\mathbf{D}]_{ii} = \frac{1}{1-(1-p_i)^c},\ i = 1\cdots, mn.$$

for $t = 1, 2, \cdots, c$ **do**
  Sample index $i_t \sim \text{categorical}(\mathbf{p})$
  $\mathcal{I} = \mathcal{I} \cup \{i_t\}$
end
Set $S_{ii} = 1, \forall i \in \mathcal{I}$

Output: $\hat{\mathbf{Z}} = \mathbf{S}\mathbf{D}\mathbf{Z}$

Algorithm 2: MVM Solver.

Solve: $\min_{\mathbf{p}\geq 0, \mathbf{1}^\top\mathbf{p}=1} \sum_{i=1}^{mn} \frac{\alpha_i}{1-(1-p_i)^c}$

Input : $[\alpha_1, \alpha_2, \ldots, \alpha_{mn}], I_{\max}, \gamma_{\text{tolerance}}$

Output: Sampling probabilities $\mathbf{p} = [p_1, p_2, \ldots, p_{mn}]^\top$

Initialize $\forall i : \mathbf{p}_i^{(1)} = \begin{cases} 0, & \text{if } \alpha_i = 0 \\ \frac{1}{\text{nnz}(\alpha)}, & \alpha_i \neq 0 \end{cases}$, and set $k = 1$

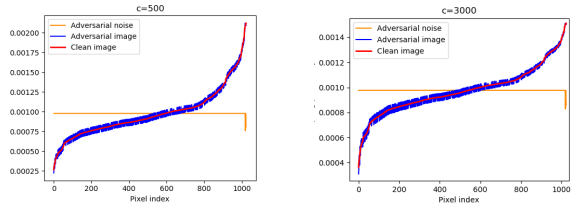
while $k < I_{max}$ *and* $\frac{\|\mathbf{p}^{(k+1)} - \mathbf{p}^{(k)}\|_2}{\|\mathbf{p}^{(k)}\|_2} > \gamma_{tolerance}$ **do**

  $r_i = p_i^{(k)} + \mu \frac{\alpha_i c (1-p_i^{(k)})^{c-1}}{\left(1-(1-p_i^{(k)})^c\right)^2} \quad \forall i$

  $p_i^{(k+1)} = \Pi_{\text{simplex}}([r_1, r_2, \cdots, r_{mn}])$

  $k = k + 1$

end



(a) (b)

Fig. 1. Optimal sampling probabilities across pixel index for matrices corresponding to noise, clean, and adversarial image. Optimal sampling probabilities for the FGSM noise $\mathbf{N}$ is uniform, while that of the clean image $\mathbf{X}_{\text{clean}}$ can be effectively estimated using $\mathbf{X}_{\text{adv}}$.

fact that noise perturbation is usually much smaller than the image $\mathbf{N} \ll \mathbf{X}_{\text{clean}}$, one can practically approximate the optimal sampling probabilities by solving (3) for $\mathbf{X}_{\text{adv}}$; that is,

$$\min_{\mathbf{p}\geq 0, \mathbf{1}^\top \mathbf{p}=1} \sum_{i=1}^{n} \frac{1}{1-(1-p_i)^c} \|\mathbf{x}_i^{\text{adv}}\|_2^2, \tag{4}$$

where $\mathbf{X}_{\text{adv}} := [\mathbf{x}_1^{\text{adv}}, ..., \mathbf{x}_{mn}^{\text{adv}}]^\top$.

To provide empirical evidence on the usefulness of approximating (3) with (4), Fig. 1 depicts the sampling probabilities across different pixels, sorted in increasing order, for a generic CIFAR sample image perturbed with adversarial FGSM noise [25]. For comparison, the optimal sampling probabilities with respect to the noise component solely is also plotted, which results in optimal uniform sampling, since for all rows of the noise component we have $\|\mathbf{n}_i\| = constant \quad \forall i$. In contrast, the solution to (4) is a practical yet accurate approximation of that in the oracle-given (3), where the small jitters are due to the adversarial noise. A similar pattern is observed across different samples, which justifies this approximation. Furthermore, we observe that as the number of draws c is increased, optimal probabilities exhibit less variability, as opposed to that for small values of c, which highlights the importance of optimal sampling when the number of draws is limited.

In a general CNN, in addition to the input, the values in mid-layers are also tensors of size $m_k \times n_k \times h_k$ for the k-th hidden layer. Thus, for a given image, the proposed approach can be readily generalized for sampling tensors at any given layer in any pre-trained CNN as well, where the probabilities are obtained per sampling unit.

2.1. Detection of adversarial images

In this subsection, we introduce how the proposed sampling units can be utilized for detection of adversarial images. The inference network is constructed by adding M number of sampling units in the trained network right after the ReLU activation units, and in the layers prior to flattening; see also Fig. 2. Then, classification of input image $\mathbf{X}_\nu$ into either of the adversarial or clean classes is carried out by: (s1) passing the image through the original inference network (no sampling) to obtain sampling probabilities per sampling unit, (s2) passing the image R number of times through the reinforced network; (s3) measuring the *uncertainty* of the network output class via entropy; and, (s4) declaring "adversary" if the entropy is higher than threshold τ, and "clean" otherwise. The algorithm is tabulated in Alg. 3 in detail, where $\mathbf{y}_\nu^{(r)}$ is the *hard* classification output for the image in the r-th pass, and Table 1 compares the sampling complexity of MVM versus stochastic approximate pruning (SAP) [24].

Algorithm 3: Detection of adversarial image.

Input: Test image $\mathbf{X}_\nu$, inference CNN reinforced with M defense units, R and τ

Pass image $\mathbf{X}_\nu$ in the full network and obtain sampling probabilities for each sampling unit accordingly

for $r = 1, 2, \cdots, R$ **do**

  Collect output class $\hat{\mathbf{y}}_\nu^{(r)}$

end

Form histogram $\{\hat{\mathbf{y}}_\nu^{(r)}\}_{r=1}^R$ and calculate entropy of $\hat{\mathbf{y}}_\nu$

Output: Declare *adversary* if entropy exceeds threshold τ

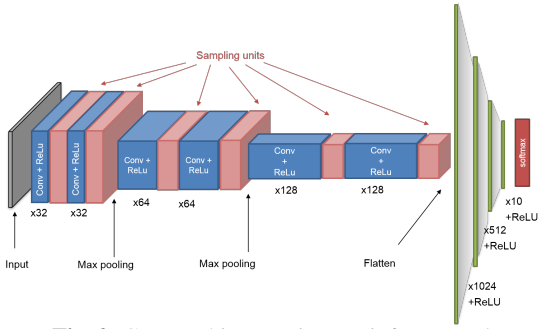


Fig. 2. CNN architecture in test (inference) phase.

	Per draw complexity	Draws needed per q elements
SAP	$\mathcal{O}(mnp)$	$\mathcal{O}(q)$
MVM	$\mathcal{O}(mn)$	$\mathcal{O}(q/p)$

Table 1: Sampling complexity for SAP and MVM schemes on a tensor of size $m \times n \times p$.

3. NUMERICAL TESTS

In this section, we test the proposed sampling scheme for detection of adversarial images for benchmark image classification datasets. To this end, we have borrowed CNN structures for classification of CIFAR-10, MNIST, and SVHN datasets along with adversarial settings from [17], and are made available online[1]. Clean and adversarial image classification accuracies against Fast Gradient Sign Method (FGSM) [25] , Jacobian Saliency-Map Attack (JSMA)[26], and Basic Iterative Method (BIM) [27] attacks, are reported in Table 2 for completeness.

	MNIST	SVHN	CIFAR-10
Clean	98.7 %	92.2 %	82.6 %
JSMA	2.70%	0.32%	0.20%
FGSM	5.87%	3.29%	7.03%
BIM-A	0.00%	0.00%	0.57%

Table 2: Classification accuracy on clean and adversarial images.

In order to properly evaluate accuracy in detection of adversarial images, we only perturb test samples that are correctly classified by the original network, since an adversary would have no incentive to perturb samples that are already misclassified. We have placed sampling units based on Dropout, SAP, and MVM schemes right

[1]https://github.com/FatemehSheikholeslami

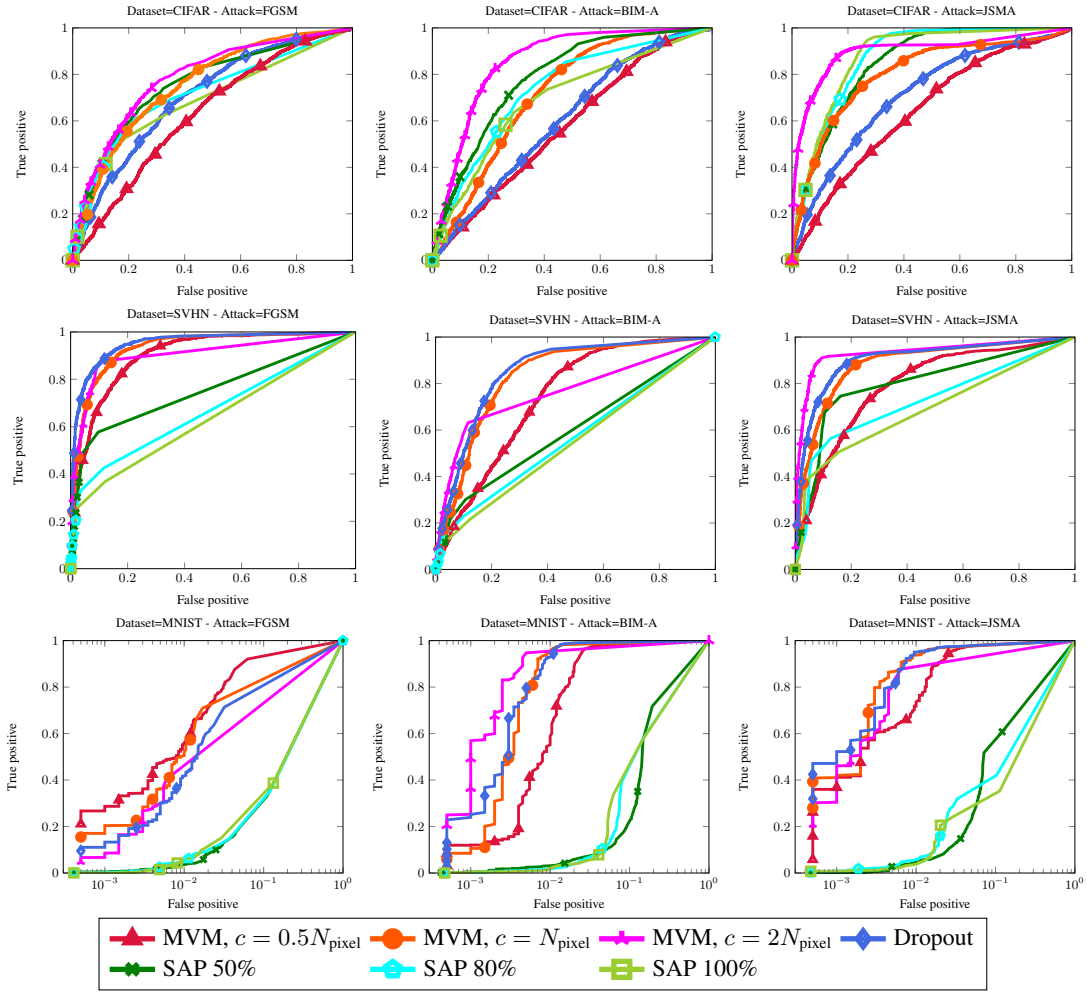


Fig. 3. ROC-curve for detection of adversarial images with different attack-detection sampling schemes.

		MNIST			SVHN			CIFAR		
		FGSM	BIM-A	JSMA	FGSM	BIM-A	JSMA	FGSM	BIM-A	JSMA
MVM	$c = 0.5N_{\text{pixel}}$	**0.947**	0.982	**0.983**	0.901	0.741	0.794	0.630	0.578	0.643
	$c = N_{\text{pixel}}$	0.848	**0.987**	0.982	**0.933**	**0.834**	0.885	0.756	0.718	0.804
	$c = 2N_{\text{pixel}}$	0.699	0.972	0.936	0.906	0.766	**0.934**	**0.787**	**0.858**	**0.899**
SAP	50%	0.613	**0.749**	**0.746**	**0.752**	**0.601**	**0.804**	**0.757**	**0.785**	0.850
	80%	0.618	0.713	0. 667	0.667	0.568	0.729	0.718	0.736	0.871
	100%	**0.631**	0.715	0.628	0.635	0.550	0.692	0.688	0.694	**0.878**
Dropout	$p_{\text{drop}} = 0.5$	**0.845**	**0.991**	**0.984**	**0.951**	**0.853**	**0.904**	**0.701**	**0.606**	**0.710**

Table 3: AUC-ROC of different attack-detection sampling schemes. Higher values indicate better detection.

after the ReLU activation units prior to flattening; e.g., see Fig. 2 for inference network used for CIFAR-10 dataset.

Fig. 5 plots the ROC curve for detection of adversarial versus clean images, where the curve is obtained by varying the threshold τ, as discussed in Alg. 3. We set $R = 100$, and the number of pixels in MNIST, SVHN, and CIFAR images are $N_{\text{pixel}} = 784, 1024$, and 1024, respectively. We tested SAP with 50%, 80% and 100% sampling ratios, and for dropout, varying the drop probability in $p_{\text{drop}} = \{0.1, 0.2, 0.5\}$ showed negligible effect in the ROC-curve, thus we are reporting the results for $p_{\text{drop}} = 0.5$. Also note that since most attack detection algorithms are very successful in the MNIST dataset, the x-axis is in logarithmic scale for better visualization.

In addition to the ROC curves, the area-under-curve (AUC) is also provided in Table 3, which further quantifies the accuracy of attack detection across different methods. Interestingly, it is observed that a smaller number of draws c provides a powerful detection in simpler images such as the MNIST dataset, while in more detailed images such as SVHN and particularly CIFAR-10, a larger number of draws yields higher AUC. Moreover, MVM-based detection can provide upto 10% improved AUC in MNIST as well as CIFAR-10 datasets with FGSM and BIM-A perturbations, respectively, while the AUC is almost the same or higher compared to the best of competing state-of-the-art detection methods in other cases. Further analysis of randomized defenses and potential improvements on larger datasets and CNNs are among future directions.

4. REFERENCES

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Intl. Conf. on Learning Representations*, San Diego, CA, USA, May 2015.

[2] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, June 2017, pp. 6517–6525.

[3] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems*, Montréal, Canada, Dec. 2014, pp. 3104–3112.

[4] J. Engel, C. Resnick, A. Roberts, S. Dieleman, M. Norouzi, D. Eck, and K. Simonyan, "Neural audio synthesis of musical notes with wavenet autoencoders," in *Intl. Conf. on Machine Learning*, Sydney, Australia, May 2017, pp. 1068–1077.

[5] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *Intl. Conf. of Learning Representations*, Banff, Canada, May 2014.

[6] S. M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, June 2017, pp. 1765–1773.

[7] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, June 2018, pp. 1625–1634.

[8] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu, "Dolphinattack: Inaudible voice commands," in *ACM Conference on Computer and Communications Security*, Dallas, TX, USA, Oct. 2017, pp. 103–117.

[9] G. Goswami, N. Ratha, A. Agarwal, R. Singh, and M. Vatsa, "Unravelling robustness of deep learning based face recognition against adversarial attacks," *arXiv preprint arXiv:1803.00401*, 2018.

[10] A. J. Bose and P. Aarabi, "Adversarial attacks on face detectors using neural net based constrained optimization," *arXiv preprint arXiv:1805.12302*, 2018.

[11] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning: From phenomena to black-box attacks using adversarial samples," *Technical Report, arXiv preprint arXiv:1605.07277*, 2016.

[12] N. Carlini and D. Wagner, "Adversarial examples are not easily detected: Bypassing ten detection methods," in *ACM Workshop on Artificial Intelligence and Security*, Dallas, TX, USA, Oct. 2017, pp. 3–14.

[13] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *IEEE Symposium on Security and Privacy*, San Jose, CA, USA, May 2017, pp. 39–57.

[14] S. Gu and L. Rigazio, "Towards deep neural network architectures robust to adversarial examples," *Intl. Conf. on Learning Representations*, San Diego, CA, USA, May 2015.

[15] J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff, "On detecting adversarial perturbations," *Intl. Conf. on Learning Representations*, Toulon, France, April 2017.

[16] J. Lu, T. Issaranon, and D. Forsyth, "Safetynet: Detecting and rejecting adversarial examples robustly," in *IEEE International Conference on Computer Vision*, Venice, Italy, Oct. 2017, pp. 446–454.

[17] R. Feinman, R. R. Curtin, S. Shintre, and A. B Gardner, "Detecting adversarial samples from artifacts," *arXiv preprint arXiv:1703.00410*, 2017.

[18] C. Guo, M. Rana, M. Cisse, and L. Van Der Maaten, "Countering adversarial images using input transformations," *Intl. Conf. on Learning Representations*, Vancouver, Canada, May 2018.

[19] T. Miyato, S. Maeda, S. Ishii, and M. Koyama, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[20] L. Schmidt, S. Santurkar, D. Tsipras, K. Talwar, and A. Madry, "Adversarially robust generalization requires more data," *Advances in Neural Information Processing Systems*, Montreal, Canada, Dec. 2018.

[21] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," *Intl. Conf. on Learning Representations*, Toulon, France, April 2017.

[22] H. Gouk, E. Frank, B. Pfahringer, and M. Cree, "Regularisation of neural networks by enforcing Lipschitz continuity," *arXiv preprint arXiv:1804.04368*, 2018.

[23] Y. Gal and Z. Ghahramani, "A theoretically grounded application of dropout in recurrent neural networks," in *Advances in Neural Information Processing Systems*, Barcelona, Spain, Dec. 2016, pp. 1019–1027.

[24] G. S. Dhillon, K. Azizzadenesheli, Z. C. Lipton, J. Bernstein, J. Kossaifi, A. Khanna, and A. Anandkumar, "Stochastic activation pruning for robust adversarial defense," *Intl. Conf. on Learning Representations*, Vancouver, Canada, April 2018.

[25] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *Intl. Conf. on Learning Representations*, San Diego, CA, USA, May 2015.

[26] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *IEEE European Symposium on Security and Privacy*, Saarbrücken, Germany, March 2016, pp. 372–387.

[27] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *Intl. Conf. on Learning Representations*, Toulon, France, April 2017.