INTELLIGENT AND ADAPTIVE MIXUP TECHNIQUE FOR ADVERSARIAL ROBUSTNESS

Akshay Agarwal[1]*, Mayank Vatsa*[2]*, Richa Singh*[2]*, and Nalini Ratha*[1]

[1]SUNY Buffalo, NY, USA and [2]IIT Jodhpur, India

ABSTRACT

Deep neural networks are generally trained using large amounts of data to achieve state-of-the-art accuracy in many possible computer vision and image analysis applications ranging from object recognition to natural language processing. It is also claimed that these networks can memorize the data which can be extracted from the network parameters such as weights and gradient information. The adversarial vulnerability of the deep networks is usually evaluated on the unseen test set of the databases. If the network is memorizing the data, then the small perturbation in the training image data should not drastically change its performance. Based on this assumption, we first evaluate the robustness of deep neural networks on small perturbations added in the training images used for learning the parameters of the network. It is observed that, even if the network has seen the images it is still vulnerable to these small perturbations. Further, we propose a novel data augmentation technique to increase the robustness of deep neural networks to such perturbations.

Index Terms— Deep Networks, Robustness, Adversarial Perturbations, Data Augmentation Technique, Object Recognition

1. INTRODUCTION

The success of deep neural networks in image recognition is partially attributed to the large databases containing millions of images. These networks are trained using millions of images and assumed to memorize the training data. The problem of memorization can be seen from the point of overfitting [1]. Such memorization or overfitting of data leads to the leakage of private data used in the training [2, 3]. While the vulnerability of deep networks against perturbation [4, 5] or image manipulation [6] have been widely explored now, the perturbation is usually added in the unseen testing images. However, in this research, we pose a question, while these deep neural networks overfitted on the training data are these networks still vulnerable against small perturbations in seen training images?

Roy et al. [6] have shown that degradations such as the incorporation of image noises can lead to drop in the recognition performance of convolutional neural networks (CNNs). Similarly, in literature, several adversarial attacks are proposed which can induce the noise to fool the CNNs without utilizing the network information [7, 8, 9, 10]. While it is observed that deep networks are highly vulnerable but it is to be noted that these images on which adversarial noise is added are unseen to the network. In this research, we have first applied the image noise on the training images itself to evaluate whether the network still robust or not. Using the experiments on multiple object recognition databases it is seen that, even the network has seen the images while learning its parameter can misclassify the same images when comes with small perturbations.

Considering the vulnerability of pre-trained CNNs using the small image perturbations and to advance the current defenses [11, 12, 13, 14], we have proposed a data augmentation technique to improve their robustness against learning-based adversarial perturbations. Fig. 1 illustrates the key idea of this research: (a) shows the traditional training of a CNN classifier and during real-world evaluated (part (b)), it can lead to an incorrect prediction on the test set because of the existence of multiple anomalies such as adversarial perturbation. Part (c) focuses on finding the singularities through seen training images but slightly distorted due to image noise and see how robust is the network even on seen images. Based on the vulnerability of the classifier on the image classes, data augmentation has been performed in (d) to retrain the sensitive model (M1). The final output is the retrained model.

Thanks to XYZ agency for funding.

Table 1: Configuration of the custom CNNs.

CNN	Configuration
Custom-1 (F-MNIST)	Conv($5 \times 5 \times 32$), ReLU, MaxPool(3×3) Conv($8 \times 8 \times 64$), ReLU, AvgPool(3×3), Conv($8 \times 8 \times 64$), ReLU, AvgPool(3×3), Fully Connected(64), ReLU, Fully Connected(10), SoftMax
Custom-2 (F-MNIST)	Conv($8 \times 8 \times 64$), ReLU, Conv($6 \times 6 \times 128$), ReLU, Conv($5 \times 5 \times 128$), ReLU, Fully Connected(10), SoftMax
Custom (CIFAR)	Conv($5 \times 5 \times 32$), ReLU, MaxPool(3×3), Conv($8 \times 8 \times 64$), ReLU, AvgPool(3×3), Conv($8 \times 8 \times 64$), ReLU, AvgPool(3×3), Fully Connected(64), Fully Connected(10), SoftMax

2. VULNERABILITIES ON SEEN IMAGES

In this section, we perform the ablation study to analyze the robustness of classifiers if they have seen the noise in the training images. The analysis corresponds to point (a) and (c) of Fig. 1. The experiments are performed using multiple databases including CIFAR10 and Imagenette, and CNN architectures including VGG and ResNet. The vulnerability analysis is conducted using the Gaussian noise, which refers to a black-box setting where no network information has been utilized to fool them. The Gaussian with mean (μ) 0 and varying standard deviation (σ) has been applied and their classification is observed against multiple databases and networks.

CIFAR10 [15] is one of the most popular object recognition databases containing low-resolution object images belonging to 10 classes. Three different CNN architectures are trained on the training set of the database namely VGG [16], ResNet [17], and Custom CNN (Table 1). The selected architectures represent the broad category of networks in terms of the number of layers and type of connections such as sequential and skip identity. Each network is trained using 'Adam' optimizer with a batch size of 32.

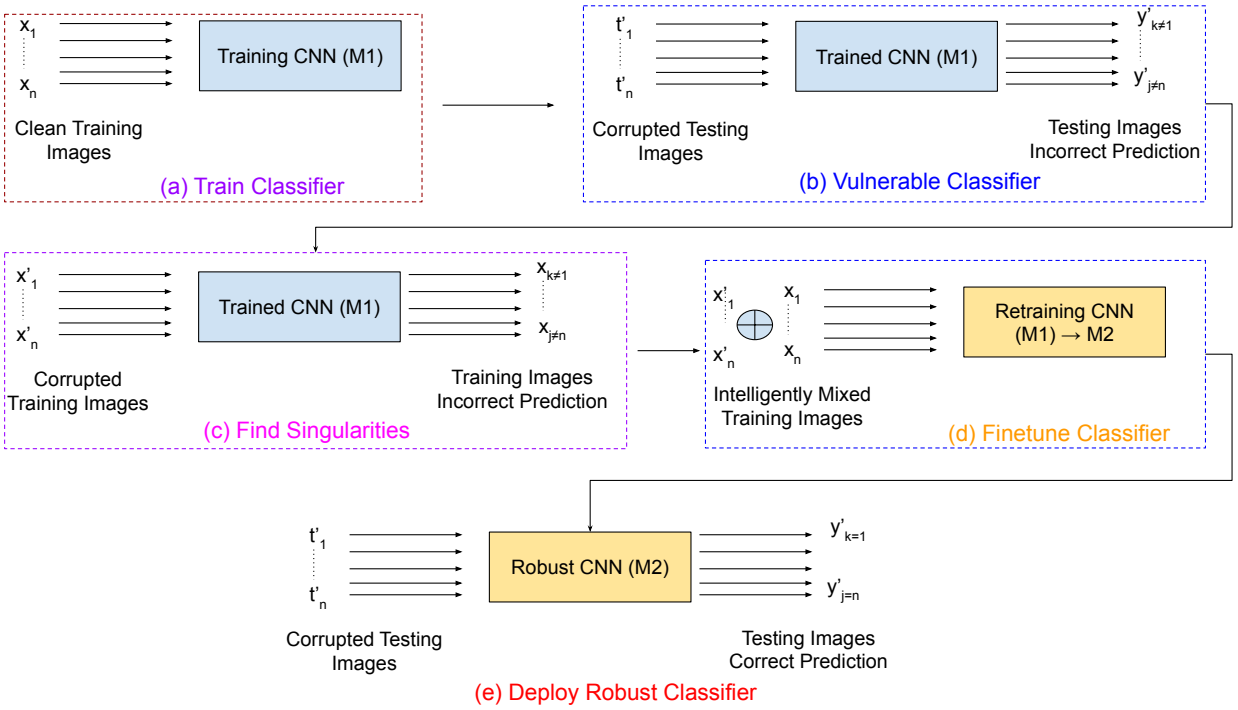

Fig. 1: Schematic diagram of the proposed research (train → find singularities → finetune for robustness).

Table 2: Sensitivity analysis (%) on CIFAR10 database using training noisy images. On the training set, the VGG, ResNet, and custom model yields 88.57%, 98.68%, and 78.45% recognition performance, respectively. $\downarrow_{-x\%}$ shows the $\sim x\%$ drop in performance when the noisy training images are given for classification.

Noise	VGG	ResNet	Custom
GN01	$74.94 \downarrow_{-13\%}$	$24.97 \downarrow_{-74\%}$	$58.63 \downarrow_{-20\%}$
GN05	$42.94 \downarrow_{-45\%}$	$14.53 \downarrow_{-84\%}$	$28.23 \downarrow_{-50\%}$
GN08	$33.59 \downarrow_{-55\%}$	$13.72 \downarrow_{-85\%}$	$22.41 \downarrow_{-56\%}$

The sensitivity analysis on the CIFAR10 database is provided in Table 2. The original models yield higher recognition performance on clean training images, however, it is found that even if the network has seen the images while training, it can still misclassify them. Out of all the networks used on CIFAR10, the ResNet model is found to be highly susceptible against noisy variation in the images. Even in the presence of as low as Gaussian noise of $\sigma = 0.01$ leads to a drop of 74%. The difference between the ResNet model from the other models is the skip connection. It seems the skip connect which helps in better gradient flow also leads to higher sensitivity.

Compared to different networks, VGG is found to be the least sensitive. The primary difference between the VGG model and the custom model apart from being the deeper layers is the size of the convolutional kernels. The VGG model uses small spatial kernels of size 3×3, whereas, in the custom model higher kernel sizes 5×5 and 8×8 are used for parameter learning. The lower kernel size is exposed to lesser pixels which might be of the same distribution, whereas, the higher kernel can lead to visualizing the different local field information. Hence, end in a higher vulnerability.

The analysis on Fashion MNIST (F-MNIST) and Imagette[1] a subset of ImageNet are reported in Fig. 2. On F-MNIST two custom models are developed (Table 1), whereas, on Imagenette, the VGG16 model has been trained. The custom models on F-MNIST yield at-least 89.76% recognition accuracy which drops down to 46% through the modification of training data through small perturbations. Interestingly, both the custom models have shown similar sensitivity in terms of the reduction of recognition performance. The prime difference between the custom models is the number of filters, their spatial dimension, pooling layers, and the use of the FC layer as a feature extractor in custom-1.

Similar to the previous two databases which are of low resolution, a network trained on a large resolution database namely ImageNette which is a subset of ImageNet [18] is found vulnerable. However, as compared to low-resolution object images of CIFAR-10, the network is found less vulnerable and need a slightly higher noise magnitude. It shows that the higher the amount of information available for feature extraction and parameter learning, the better the robustness of the network. The VGG-16 network shows more than 97% recognition performance on the training set, which drops to 56.95% when Gaussian noise with $\sigma = 0.08$ is used. The relative reduction in the performance of VGG on CIFAR-10 and ImageNette are $\sim 55\%$ and $\sim 40\%$ on GN08, respectively.

We believe this information can help the research community in developing the classification network by being aware of the vulnerability of the individual constituents such as filter size and type of connection, (e.g. skip connection, pooling layer, fully connected layer).

[1] https://github.com/fastai/imagenette

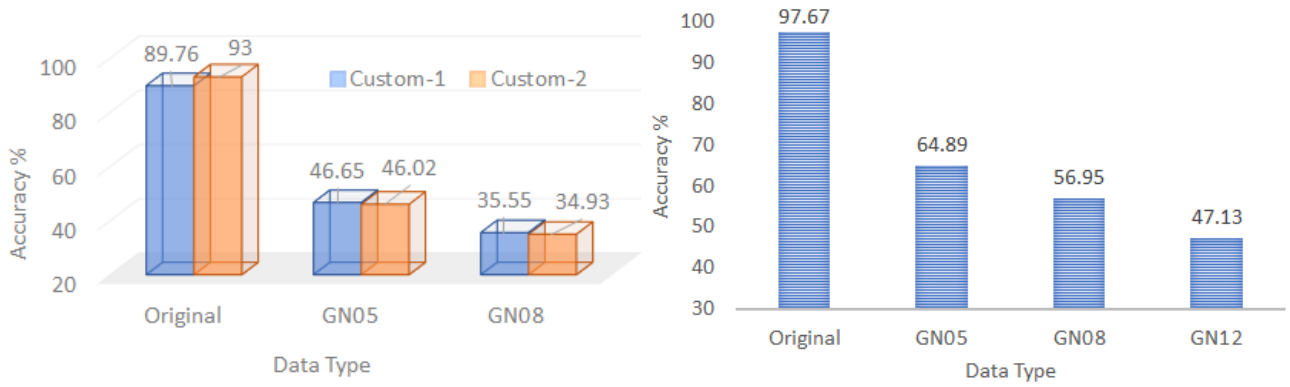

Fig. 2: Sensitivity analysis on Fashion-MNIST and Imagenette database using the recognition performance on training noisy images. On both databases, each network including shallow and deep networks shows a significant reduction in performance.

3. PROPOSED DEFENSE FOR DEPLOYABLE ROBUST MODEL

Through the analysis in the previous section, we have proposed a novel technique of data augmentation for possible adversarial robustness. The proposed technique is inspired by the algorithm termed as 'mixup' [19]. The mixup algorithm combines the two data points and corresponding labels through the mixing coefficients. The data points are randomly selected from the training images of the database. The intuition of using the mixup is to reduce the drawbacks of memorization of data, adversarial examples, and promote linear behavior among the data points. Mathematically, the mixup algorithm can be written as follows:

$$\begin{aligned} \tilde{x} &= \lambda x_1 + (1-\lambda)x_2 \\ \tilde{y} &= \lambda y_1 + (1-\lambda)y_2 \end{aligned} \tag{1}$$

where, λ is the mixup coefficient derived from the 'Beta' distribution with parameter (α, α). $\tilde{x}$ is the mixup data point and $\tilde{y}$ is the corresponding class label of the point. For merging the class labels in this fashion, the labels are first converted into one-hot encoding. x_1 and x_2 are the two data points randomly selected from the training images and y_1 and y_2 are their corresponding original labels. The primary drawback of the above mixup algorithm is that first the data points are randomly selected and hence can be of the same class. Secondly, it does not utilize any class information, in which a noisy variant of an image can be misclassified. To counter these limitations, we have proposed an extension of this algorithm by utilizing the class information in which a noisy variant of an image can be misclassified. Mathematically, the proposed **'intelligent'** mixup can be written as:

$$\begin{aligned} \tilde{x} &= \lambda x_1 + (1-\lambda)x_1' \\ \tilde{y} &= \lambda y_1 + (1-\lambda)y_1' \end{aligned} \tag{2}$$

where, x_1' is the noisy variant of x_1. y_1 is the original class label of x_1 and y_1' is the class label of x_1' predicted by the classifier. We hypothesize that the classes that generally lie close to each other in the feature space get highly misclassified as compared to the ones which lie far from each other. Apart from getting advantages of traditional 'mixup', this label-aware mixup can help in learning the generalized features grouping similar classes close to each other and different classes lying far in the feature space. We have experimentally found and used $\alpha = 8.0$ for mixing up the data points.

Table 3: Adversarial robustness (%) of VGG network on CIFAR-10 dataset.

Attack	Undefended	Proposed Defense with	
		GN05	GN08
CW	10.20	**74.65**	71.60
DeepFool	10.85	**33.50**	30.80
FGSM	18.35	**62.00**	61.75
40-IFGSM	16.25	59.20	**60.50**
100-PGD	16.10	61.55	**61.85**
No attack (clean)	83.91	**85.20**	83.15

4. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we first describe the ingredients used to perform the experiments such as database, attacks, and classifiers. Later, the results corresponding to vulnerable and defended networks using the proposed '*intelligent*' mixup generation are provided.

4.1. Experimental Setup

In this research, we have used Fashion-MNIST (F-MNIST) [20] and CIFAR-10 databases. These datasets are popular object recognition databases containing 50k and 60k training images, respectively. The noisy variant of the clean training images of each database is generated using two standard deviations of Gaussian noise such as $\sigma = 0.05$ and $\sigma = 0.08$. The noisy variant is referred to as GN05 and GN08 according to the σ used for manipulation.

We have used VGG-16 network [16] for recognition on CIFAR-10 dataset, while the custom-2 model described in Table 1 is used for F-MNIST dataset. The networks are pretrained using Adam optimizer, the batch size is set to 32, and the learning rate is set adaptively with an initial value of 0.0001. Once the noisy variants are obtained, they are used for classification using the trained classifiers, and the label given by the classifier on the noisy variant is used for the proposed 'intelligent' mixup. After the generation of mixup images, they are augmented to the original training set of the database and used for finetuning the adversarially vulnerable model for 30 epochs.

Multiple complex adversarial attack algorithms are used for generating the adversarial examples varying from optimization based at-

Table 4: Adversarial robustness (%) of custom-2 model on F-MNIST.

Attack	Undefended	Proposed Defense with	
		GN05	**GN08**
CW	6.45	71.54	**72.87**
DeepFool	12.09	33.73	**33.76**
FGSM	9.85	26.20	**26.51**
40-IFGSM	11.41	**31.57**	28.85
100-PGD	3.01	**21.51**	13.22
No attack (clean)	91.49	90.34	**90.46**

tacks such as C&W l_2 [21] and DeepFool [22] and gradient-based attacks such as single step FGSM [23], k_1-FGSM [24], and k_2-PGD [25]. The standard attack strength parameter is used on each database, i.e., $\epsilon = 0.03$ is used on the color images and $\epsilon = 0.3$ on gray-scale images. For iterative FGSM attacks 40 steps (i.e., k_1=40) and for iterative PGD 100 steps (i.e., k_2=100) are used for adversarial examples. We have used the standard attacks setting unless otherwise specified.

4.2. Results and Observations

The adversarial robustness performance on CIFAR-10 is reported in Table 3. Adversarial robustness refers to the improvement in the classification accuracy on the adversarial examples. From the point of adversarial attacks, the network is found to be highly vulnerable against optimization-based attacks as compared to both single and multi-step gradient based attacks. The CW l_2 optimization attack which reduces the performance of the VGG network to 10.20% is found to be least effective when the proposed defense is applied on the VGG network. Apart from that, it is found in the literature while the adversarial defense algorithms can provide robustness to the networks on adversarial examples but significantly reduces the performance on clean images. However, the proposed defense is not only able to increase the robustness of the network, it is able to retain the performance on clean images or improve it.

In contrast to the findings on color object images of CIFAR-10 dataset, on grayscale images the proposed defense, with higher σ, yields better robustness on optimization attacks such as CW l_2 and DeepFool. While on iterative gradient variants, augmentation with lower σ shows higher robustness. However, in both the cases, the proposed defense is found significantly effective in handling adversarial images. The adversarial vulnerability and robustness results are corresponding to the point (b) and (e) in Fig. 1, where the undefended and final robust model is used for evaluation, respectively.

5. CONCLUSION

In the context of increasing the robustness of the CNN classifiers against adversarial examples, this paper presents two key contributions. First, we have experimentally shown that the networks that have even seen the images at the time of training are vulnerable when small perturbation is added. It is to be noted that this noise may not utilize any network information which adversarial attacks typically utilize while crafting the perturbation. In the experiments, we have observed that even black-box noise can significantly reduce the performance of deep networks. It is also found that the skip connection-based ResNet architecture posses higher vulnerability as compared to a sequentially connected network. Next, by utilizing the incorrectly predicted label of the networks on the noisy variant of the training images, a novel '*intelligent*' mixup approach is proposed. The proposed mixup based data augmentation technique can increase the adversarial robustness of the undefended networks trained using clean images only.

Acknowledgements

M. Vatsa is partially supported by the Department of Science and Technology, Government of India through Swarnajayanti Fellowship.

6. REFERENCES

[1] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals, "Understanding deep learning requires rethinking generalization," *arXiv preprint arXiv:1611.03530*, 2016.

[2] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al., "Extracting training data from large language models," *arXiv preprint arXiv:2012.07805*, 2020.

[3] Milad Nasr, Reza Shokri, and Amir Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in *IEEE S&P*, 2019, pp. 739–753.

[4] Alex Serban, Erik Poll, and Joost Visser, "Adversarial examples on object recognition: A comprehensive survey," *ACM CSUR*, vol. 53, no. 3, pp. 1–38, 2020.

[5] Richa Singh, Akshay Agarwal, Maneet Singh, Shruti Nagpal, and Mayank Vatsa, "On the robustness of face recognition algorithms against attacks and bias," *AAAI*, pp. 13583–13589, 2020.

[6] Prasun Roy, Subhankar Ghosh, Saumik Bhattacharya, and Umapada Pal, "Effects of degradations on deep neural network architectures," *arXiv preprint arXiv:1807.10108*, 2018.

[7] Gaurav Goswami, Nalini Ratha, Akshay Agarwal, Richa Singh, and Mayank Vatsa, "Unravelling robustness of deep learning based face recognition against adversarial attacks," in *AAAI*, 2018, pp. 6829–6836.

[8] Gaurav Goswami, Akshay Agarwal, Nalini Ratha, Richa Singh, and Mayank Vatsa, "Detecting and mitigating adversarial perturbations for robust face recognition," *IJCV*, vol. 127, no. 6-7, pp. 719–742, 2019.

[9] Akshay Agarwal, Mayank Vatsa, Richa Singh, and Nalini K Ratha, "Noise is inside me! generating adversarial perturbations with noise derived from natural filters," in *IEEE/CVF CVPRW*, 2020, pp. 3354–3363.

[10] Akshay Agarwal, Richa Singh, and Mayank Vatsa, "The role of 'sign' and 'direction' of gradient on the performance of CNN," in *IEEE/CVF CVPRW*, 2020, pp. 2748–2754.

[11] Akshay Agarwal, Richa Singh, Mayank Vatsa, and Nalini K Ratha, "Image transformation based defense against adversarial perturbation on deep learning models," *IEEE TDSC*, 2020.

[12] Akshay Agarwal, Gaurav Goswami, Mayank Vatsa, Richa Singh, and Nalini K Ratha, "DAMAD: Database, attack, and model agnostic adversarial perturbation detector," *IEEE TNNLS*, 2021.

[13] Saheb Chhabra, Akshay Agarwal, Richa Singh, and Mayank Vatsa, "Attack agnostic adversarial defense via visual imperceptible bound," in *IEEE ICPR*, 2021, pp. 5302–5309.

[14] Akshay Agarwal, Mayank Vatsa, and Richa Singh, "Role of optimizer on network fine-tuning for adversarial robustness (student abstract)," in *AAAI*, 2021, pp. 15745–15746.

[15] Alex Krizhevsky and Geoffrey Hinton, "Learning multiple layers of features from tiny images," *Thesis, University of Toronto*, pp. 32–35, 2009.

[16] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *ICLR*, 2015.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *IEEE CVPR*, 2016, pp. 770–778.

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE CVPR*, 2009, pp. 248–255.

[19] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," in *ICLR*, 2018.

[20] Han Xiao, Kashif Rasul, and Roland Vollgraf, "Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[21] Nicholas Carlini and David Wagner, "Towards evaluating the robustness of neural networks," in *IEEE S&P*, 2017, pp. 39–57.

[22] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *IEEE CVPR*, 2016, pp. 2574–2582.

[23] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples," *ICLR*, 2015.

[24] Alexey Kurakin, Ian Goodfellow, and Samy Bengio, "Adversarial examples in the physical world," *arXiv preprint arXiv:1607.02533*, 2016.

[25] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, "Towards deep learning models resistant to adversarial attacks," in *ICLR*, 2018.