Bolstering Adversarial Robustness with Latent Disparity Regularization

David Schwartz and Gregory Ditzler
Department of Electrical and Computer Engineering
University of Arizona
Tucson, United States of America
dmschwar@email.arizona.edu, ditzler@arizona.edu

***Abstract*—Recent research has revealed that neural networks and other machine learning models are vulnerable to adversarial attacks that aim to subvert their predictions' integrity or privacy by adding a small calculated perturbation to inputs. Further, the adversary can significantly degrade the performance of the model. The number and severity of attacks continues to grow. However, a dearth of techniques robustly defends machine learning models in a computationally inexpensive way. Against this background, we propose an adversarially robust training procedure and objective function for arbitrary neural network architectures. Robustness of neural networks against adversarial attacks on integrity is achieved by augmentation of a novel regularization term. This regularizer penalizes the discrepancy between the representations induced in hidden layers by benign and adversarial data. We benchmark our regularization approach on the Fashion-Mnist and Cifar-10 datasets. Our model is benchmarked against three state-of-the-art defense methods, namely: (i) regularization to the largest eigenvalue in the Fisher information matrix of the activity of the terminal layer, (ii) a higher-level representation guided denoising autoencoder (trained with adversarial examples), and (iii) training an otherwise undefended model on data distorted by additive white Gaussian noise. Our experiments show that the proposed regularizer provides significant improvements in adversarial robustness over both an undefended baseline model as well as the same model defended with other techniques. This result is observed over several adversarial budgets with only a small (but seemingly unavoidable) decline in benign test accuracy.**

***Index Terms*—Adversarial Defenses, Fast Gradient Sign Method, Regularization.**

I. Introduction

Deep neural networks have quickly become the de-facto benchmark for many challenging tasks such as image segmentation [1], image classification [2], cancer detection [3], [4], and automatic speech recognition [5]. Deep neural networks have even out-performed humans in tasks once considered intractable for machine intelligence, such as Go [6]. Unfortunately, recent research has revealed that neural networks and other machine learning models are vulnerable to adversarial attacks that aim to subvert the integrity of their output decision by adding a minute quantity of judiciously chosen noise to input samples [7], [8]. The severity of consequences of the exploitation of this weakness swells as the ubiquity of machine learning systems broadens. While the research community has developed several methods by which users of machine learning can defend neural networks from such attacks, techniques that ameliorate this risk are scarce. To further complicate matters, in some situations, there is an intrinsic and dataset-dependent trade-off between adversarial robustness, robustness in the presence of random noise, and generalization performance on benign data [9]–[12].

Specifically, adversarial robustness refers to the extent to which a neural network maintains prediction accuracy on par with accuracy achieved on noiseless (i.e., benign) data when its inputs are corrupted by adversarial noise. Such adversarial noise can, in general, be applied to input data with a deleterious effect on integrity at any phase of a neural network's existence. For example, so called 'poisoning attacks' are applied at the training phase and have been shown to degrade a neural network's performance similarly to evasive attacks, in which the adversary's influence is applied to test data [13], [14]. The existence of these vulnerabilities abruptly provoked an ongoing arms race centered between adversarial attack methodologies and techniques that defend otherwise vulnerable machine learning systems [7].

Prior investigations of a multitude of defense strategies have produced a variety of conflicting results. One such contentious technique, known as a Higher Level Representation Guided Denoiser (HGD), aims to thwart the adversary by transforming input samples before they reach the neural network under attack [15]. However, [9] demonstrated that seven of the techniques described in [7] (including HGD) were published with results that could not be replicated. Furthermore, HGD and several others among these methods fail to bestow robustness to additive white Gaussian noise (AGN), while classifiers trained on AGN-distorted examples achieved significant gains in robustness to AGN and adversarial evasion attacks [9]. Another technique was developed to impart improved adversarial robustness upon neural networks trained only on benign data by regularizing the training objective in proportion to the largest eigenvalue of the Fisher information matrix (FIM) of the activity of the output layer. In this work, we refer to this regularization technique as "FIMR" and the associated Fisher information matrix as "FIM." Other state-of-the-art defense tactics reviewed in [7] rely on invariance of hidden activities under quantization (which, in some cases, fails to hold for adversarial examples), attention mechanisms, or the inclusion of an additional ad-hoc classifier. One particularly noteworthy approach was proposed by Zheng et al. [16]. Their work

demonstrated robustness can be improved through training on additive white Gaussian noise (but only to this source of noise and not specifically to adversarially optimized noise) by penalizing the disparity that would be induced in the final (i.e., decision) layer by the original sample and a randomly corrupted version.

The primary contribution of this work is a training technique that effectively and computationally inexpensively defends neural network classifiers from evasive attacks. We propose a method that bolsters adversarial robustness with the addition of a novel regularization term, which penalizes the training objective in proportion to the discrepancy between the representations induced in hidden layers by benign and adversarial data. We demonstrate that our approach more effectively defends neural networks from adversarial evasion attacks than other state-of-the-art techniques. Moreover, our method achieves robustness to AGN comparable to the technique introduced in [9]. While they both require computation of adversarially perturbed versions of training samples, in contrast to the HGD, our training scheme only requires processing of the model to be defended, while HGD demands this in conjuction with an auxiliary neural network responsible for minimizing adversarial influence on the input to the defended model. Our approach differs from that of [16] in two important ways: our regularizer penalizes absolute differences between activity induced in *every* hidden layer by *adversarial examples* and their benign counterparts, while Zheng et al.'s regularizes in proportion to the empirical KL-divergence between activities induced by benign and *AGN-corrupted* samples solely in the *output layer*. Consequently, to some extent, our contribution is an extension and dialectical synthesis of the defense strategies put forth in [15] and [16], to depend on the activity induced in each hidden layer by adversarial and benign examples.

II. Related Work

A. Vulnerable Deep Neural Networks

Convolutional neural networks (CNNs) are a type of neural network that consists of serially connected banks of neurons, whose output states result from the convolution of an input signal with a learned function. VGG is a remarkably deep CNN that achieves near state-of-the-art performance on various complicated image classification tasks [17]–[19]. The VGG architecture consists of feed-forward banks of two-dimensional convolutional layers whose outputs propagate through max-pooling layers that operate by down-sampling their input, aggregating pixel values by passing the largest among each (2,2)-pixel wide neighborhood in the input. VGG's structure encourages a re-encoding of pixel information into spatial maps of features situated at levels of abstraction that increase as they propagate deeper in the network. In our experiments and those conducted by others, VGG has been demonstrated to be severely vulnerable to adversarial evasion attacks [20], incurring a significant breakdown in accuracy even for smaller adversarial perturbations.

B. Adversarial Machine Learning

We begin by assuming that the adversary has knowledge about the loss function, $\mathcal{L}$ [21]. This loss function is used to train a neural network with some parameters, $\boldsymbol{\Theta}$, that were learned from the dataset defined by pairs of feature vectors with labels, $(\mathbf{x}, y)$. The goal of the adversary is to generate a perturbation vector $\boldsymbol{\eta}_\epsilon$ that maximizes the likelihood of fooling the target network (e.g., lead to an incorrect classification with high confidence). It is worth noting that this perturbation vector is not random, but constructed in a way that achieves the adversary's goal of triggering a misclassification. The Fast Graident Sign Method (FGSM) generates this noise by choosing $\boldsymbol{\eta}_\epsilon$ to be

$$\boldsymbol{\eta}_\epsilon = \epsilon \text{sign}\left(\nabla_{\mathbf{x}} \mathcal{L}(\boldsymbol{\Theta}, \mathbf{x}, y)\right) \tag{1}$$

where ϵ is a small positive number controlling the strength of the attack, $\nabla_{\mathbf{x}}$ represents the gradient operation with respect to the input features and $\boldsymbol{\Theta}$ represents the parameters of the neural network. That is, $\boldsymbol{\eta}_\epsilon$ perturbs the benign sample a distance ϵ in the direction that maximizes the targeted model's loss to produce from the benign example, $\mathbf{x}$, the adversarial sample, $\mathbf{x}_{\text{a}} = \mathbf{x} + \boldsymbol{\eta}_\epsilon$. Larger values of ϵ are more likely to result in misclassifications of the given sample and ϵ can be thought of as an adversary's budget in the sense that larger perturbations tend to be more easily detected, but more likely to fool the target [22].

C. Higher Level Representation Guided Denoising Autoencoders

The computational ease with which adversarial examples can be calculated demonstrates an urgent need for training schemes and methods that render neural networks more robust to adversarial attacks. The appendage of a HGD, instantiated at the input of an inchoate neural network to transform inputs as a means to remove the adversarial influence operates so that the target neural network need not be optimized further [15]. As a result, the HGD is a desirable solution. However, its inclusion significantly increases the number of parameters that must be loaded to predict with the HGD-defended neural network. The HGD is based on a denoising convolutional autoencoder that is trained with benign samples in conjunction with their adversarially perturbed analogs [15]. This auxiliary network estimates a defensive counter-perturbation that when added to the adversarial input minimizes the disparity between hidden representations induced by the benign and adversarial samples at the penultimate layer of the target network. Hence, HGD is trained to output a defensive perturbation that annihilates the adversary's influence on the representation induced by the given sample in layer p, minimizing $|\mathbf{s}_p(\mathbf{x}) - \mathbf{s}_p(\mathbf{x}_{\text{a}})|$, where $\mathbf{s}_p(\cdot)$ is the activity induced in the penultimate layer.

D. Eigen-decomposition of the Fisher information Matrix

Another highly desirable category of adversarial defense solutions requires no adversarially generated information. The FIMR, introduced in [23], augments a neural network's loss function with a term that penalizes learning approximately in

proportion to the largest eigenvalue of an approximation of the Fisher Information Matrix (FIM) of the post-activation output of the penultimate layer of the underlying target model.

$$\mathrm{KL}(P(\mathbf{y}|\mathbf{x})\|P(\mathbf{y}|\mathbf{x}+\boldsymbol{\eta})) \propto \boldsymbol{\eta}^{\mathsf{T}}\mathbf{F}\boldsymbol{\eta}$$

where $\mathrm{KL}(p\|q)$ is the KL-divergence between probability distributions p and q, $\mathbf{F}$ is the Fisher Information Matrix [23], [24] and $P(\mathbf{y}|\mathbf{x})$ is the posterior probability produced by the model. The regularizer on the FIM penalizes against bias in confidence in the output layer, this tactic has the effect of smoothing and broadening the spectrum of the output activity, which may lower the capacity for adversarial perturbations to propagate through the final layer. As in [23], we approximate the trace of the FIM as $\sum_j \frac{1}{p_j}$, where p_j is the neural network's estimated probability that the input belongs in class j . The resulting loss function is formulated as

$$\mathcal{L} = (1-\zeta)\,\mathcal{L}_{\text{CE}} + \zeta \sum_{\text{classes } j} \frac{1}{p_j}, \tag{2}$$

where $\mathcal{L}_{\text{CE}}$ is cross entropy loss and $\zeta \in [0,1]$ is a penalty coefficient that weights the convex combination to provide a trade-off between the cross-entropy and FIM regularization.

E. Natural Noise

Gilmer et al. elucidate that as a consequence of the geometry of the image space perceived by a neural network and the latent spaces to which it projects information in its hidden layers, robustness to natural noise implies a limited robustness to adversarial noise [9]. This observation corroborates the finding of Zantedeschi et al. that noisy training (i.e., training on inputs subjected to AGN) is an effective defense strategy [25]. Additionally, it was demonstrated that a plethora of defense strategies (including HGD) that rely on information from attacked samples often fail to translate adversarial robustness to the more general setting of AGN [9]. For convenience, we refer to the noise itself as "AGN" and the corresponding defense tactic of training only on AGN-distorted examples as "AGNT."

III. Our Contribution

The related work has shown there is value in adding regularization terms into neural networks' cost functions to bolster robustness. In light of this potential to defend a neural network's performance in the face of adversarial attacks, we propose the addition of a novel regularization term proportional to the discrepancy between the representations induced in hidden layers by benign and adversarial data. Motivating this regularization term is the expectation that in an adversarially robust CNN, the discrepancy between hidden (i.e., latent) activity induced by adversarial and benign samples should be small. Therefore, we extend the cost function to penalize against large discrepancies between these latent representations. Specifically, we consider the novel loss function,

$$\mathcal{L} = (1-\xi)\,\mathcal{L}_{\text{CE}} + \frac{\xi}{\sum_{\text{layers } i} |\mathbf{s}_i(\mathbf{x})|} \sum_{\text{layers } i} |\mathbf{s}_i(\mathbf{x}) - \mathbf{s}_i(\mathbf{x}_a)| \tag{3}$$

where $\mathcal{L}_{\text{CE}}$ is the original undefended categorical cross-entropy loss function, $\xi \in [0,1]$ is a penalty coefficient, $\mathbf{s}_i(\mathbf{x})$ and $\mathbf{s}_i(\mathbf{x}_a)$ represent the activity of the i^{th} layer and the activity that would be induced in the i^{th} layer by stimulating the given neural network with an adversarially perturbed version of the same input. An essential distinction between (3) and the approach to general robustness (i.e., in the face of all noise) of [16] is immediately apparent in that our regularizer is a function of latent disparity induced in *every hidden layer*. In contrast, the work presented in [16] seeks solely to stabilize the activity of the output layer. Moreover, our regularizer penalizes adversarially induced latent disparity. Specifically, the work of Zheng et al. [16] does not train against samples that are adversarially optimized. Rather they opt to regularize to final layer's disparities induced by additive white Gaussian noise (AGN). Figure 1 exposes a graphical portrayal of signals' flow in our final training procedure, which we dub "HLDR".

IV. Experimental Framework

We evaluate the benign and adversarial test accuracy after fine-tuning in a stratified ten-fold cross-validation (CV) experiment. The benchmarks were performed in a way that reproducibly disentangles apparent trends from the non-determinism in the processes that produce the datasets considered and what is inherent in the low-level implementations of popular machine learning platforms (e.g., Tensorflow, pyTorch [26], [27]). The Fashion-Mnist and Cifar-10 datasets were selected for the experiments based on their popularity in adversarial machine learning [28], [29]. A VGG network is pre-trained on 10,000 randomly selected samples for up 1000 epochs or convergence (whichever comes first). This pre-training step halts early if the change in validation loss stagnates for 100 epochs. The pre-training data are discarded for the remainder of the experiment. The unseen training and test data are split over ten stratified subsets. For each fold of CV, 6000 adversarial examples are generated using the FGSM (i.e., see (1)) for several small budgets, which are presented as training data to the HGD and HLDR with their benign counterparts. Meanwhile, as in [23], Shen et al.'s model is trained using only the benign samples. AGNT (and other models exposed to AGN-perturbed examples) are fine-tuned on AGN-perturbed versions of the benign analogs of the aforementioned 6000 adversarial samples. In every iteration of the CV experiment, each model is evaluated on the test subset assigned to that iteration, subject to a varying adversarial budget.

A fine-tuning data subset is an ordered product of the sets of adversarially perturbed samples and their benign presentation (e.g., $(\mathbf{x}, \mathbf{x}_a)$). The adversarial samples are generated with pre-specified budgets $\frac{16}{255}$ and $\frac{32}{255}$. We define $\mathcal{D}$ and dataset $\mathcal{D}_\epsilon$ to be the benign component of the fine-tuning subset and its adversarially perturbed form with a budget of ϵ, respectively. The fine-tuning subset is

$$\mathbb{D} = \left\{ (\mathbf{x}, \mathbf{x}_a) | \mathbf{x} \in \mathcal{D}, \mathbf{x}_a = \mathbf{x} + \boldsymbol{\eta}_\epsilon, \epsilon \in \left\{ 0, \frac{16}{255}, \frac{32}{255} \right\} \right\}.$$

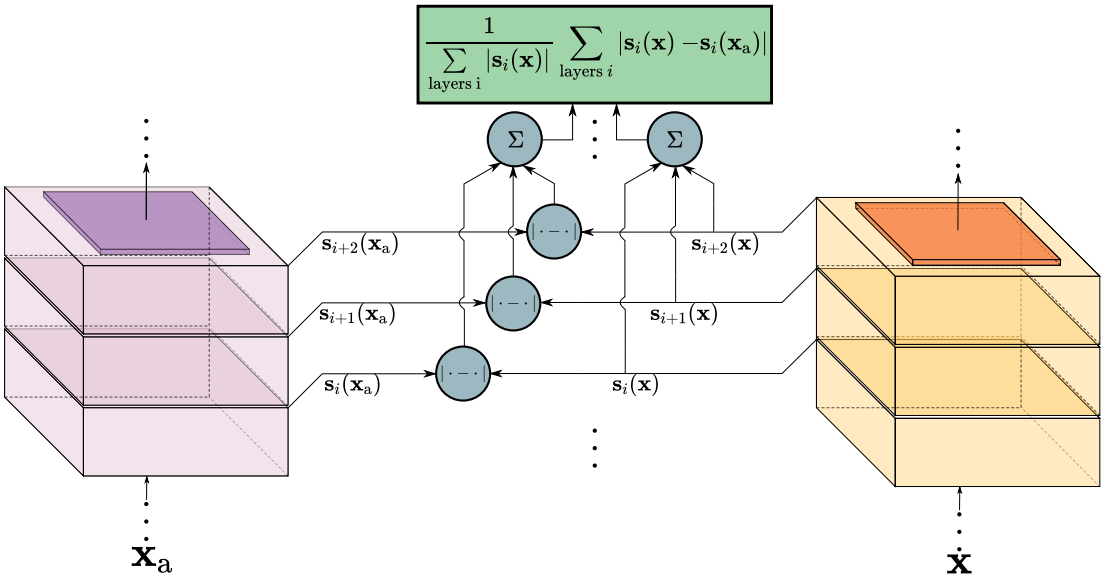


Fig. 1: An illustration of the calculation of our adversarial regularizer from three intermediate layers of VGG. The purple (left) and yellow (right) convolutional layers share weights, and are stimulated by adversarially perturbed data and benign data, respectively.

We also measure empirical accuracy in the presence of AGN (as opposed to adversarial noise).

A natural question arises from Gilmer's conclusions [9], that limits on adversarial robustness are intimately related to testing error in the presence of AGN: to what extent (and in what direction) does AGNT alter adversarial robustness and robustness to AGN? To better understand this connection, we also engineer a dual of the previously outlined experiment. In addition to training on $\mathbb{D}$, each approach is trained on examples distorted by pixelwise AGN. In this second experiment, HLDR and HGD are trained to minimize disparities between representations induced by benign and noisy pairs of samples. More precisely, we construct these extended training subsets as

$$\mathbb{D}_{\mathcal{N}} = \mathbb{D} \cup \left\{ (\mathbf{x}, \mathbf{x}_n) \,|\mathbf{x} \in \mathcal{D}, \mathbf{x}_n = \mathbf{x} + \mathbf{n}, \mathbf{n} \sim \mathcal{N}\left(0, \frac{32}{255}\right) \right\}.$$

V. Model Details

This section discusses implementation details of components of the experimental framework and their interactions with the relevant training procedures. Every fold of the CV experiment begins with the same pre-trained VGG model and is subjected to fine-tuning on $\mathbb{D}$, constructed as described in Section IV. Implementation of all models was composed in Tensorflow so that all models could be evaluated on the same platform and experimental controls. Tensorflow is initialized following the suggestions of [26], which reduces the impact of non-deterministic processes on the compilation and execution of backpropagation. Optimization of each model is performed using the "Nadam" algorithm [30], an extension of the popular "adam" routine that incorporates Nesterov momentum to increase the rate at which the optimization process converges. We implemented FIMR by incorporating the approximation evaluated in [23]. The penalty coefficient of HLDR and FIMR, respectively, are fixed as $\xi = 0.25$ and $\zeta = 0.0025$. The results presented in this work use a Feature Guided Denoiser, which is the best performing HGD introduced in [15]. This denoising autoencoder is trained to minimize $|\mathbf{s}_f(\mathbf{x}) - \mathbf{s}_f(\mathbf{x}_\text{a})|$, where $\mathbf{s}_f(\cdot)$ is the output of the final convolutional layer. We experimented with the other HGD objectives (e.g., minimizing $|\mathbf{s}_{-1}(\mathbf{x}) - \mathbf{s}_{-1}(\mathbf{x}_\text{a})|$, where $\mathbf{s}_{-1}$ represents the penultimate densely connected layer of VGG) and observed no apparent difference in benign or adversarial test accuracy.

Our VGG model is slightly different from the architecture developed by Simonyan et al. [17]. The VGG model used in these experiments is augmented with batch-normalization following every convolutional layer [31]. This implementation of VGG forgoes "relu" activation functions in favor of rectified exponential nonlinearities. The primary two blocks of our implementation of VGG have two serially connected two-dimensional convolutional layers that output 64 channels through a max-pooling operation. This pattern repeats in subsequent blocks, doubling the number of channels output with each successive block terminated by a max-pooling layer until the final two, which output 512 channels. Blocks 3-5 each have three serially connected convolutional layers before their max-pooling layers. Block 5 broadcasts into a fully connected layer of 512 neurons, re-encoding the data in its 512 input channels before feeding this information to the final soft-max layer of output neurons.

VI. Results

This section presents the empirical results from the adversarial defenses discussed throughout this manuscript. Figure 2a shows the comparisons between empirical accuracy on adversarial and benign copies of the test data for varying adversarial

budgets, ϵ, (note $\epsilon = 0$ here refers to the benign accuracy). Figure 2b show these results as the differences relative to those of the undefended model. The error bars in each figure represent 95% confidence intervals. Most surprisingly, while it retained the initially high benign test accuracy achieved by the undefended model, HGD failed to convey a significant improvement on the adversarial test set for any positive budget considered. Shen et al.'s regularizer incurs a small weakening of benign test accuracy (and an insignificant reduction in adversarial test accuracy for small ϵ) but gains a small but insignificant improvement in adversarial test accuracy for larger ϵ. Our proposed regularizer, HLDR, incurs a slight reduction in benign test accuracy and attains a remarkably large (relative to the other methods considered) improvement in adversarial test accuracy for all non-zero budgets tested.

Figures 2c and 2d show results of the same experiment applied to the Cifar-10 dataset. Notably, the phenomenon observed in the Fashion-Mnist results is present in the Cifar-10 results as well, which is that HLDR achieves significantly more adversarial robustness than FIMR and HGD. Surprisingly, as with Fashion-Mnist, HGD failed to achieve a significant improvement over the undefended model while FIMR develops a significantly more robust network (as compared to the undefended and FIMR models evaluated on Fashion-Mnist).

Figures 3a-3d depict the influence of AGN of various intensities (i.e., standard deviations, σ) on the classification accuracy of the instance of VGG protected by the corresponding defense technique and the relative change in accuracy with respect to the undefended model. Interestingly, when assessed over the Fashion-Mnist dataset (Figures 3a and 3b), the HGD-defended model is not meaningfully impacted by the random corruption applied to its input data. In contrast, the performance of the FIMR- and HLDR-defended models were observed to be more prone to a degradation in accuracy. In the case of the Fashion-Mnist dataset, neither the FIM-defended model nor HLDR were likely to resist this degradation of performance in the face of AGN. However, each model was infrequently (i.e., on few iterations of CV) observed to enjoy a small boost in classification accuracy for non-zero σ. Strikingly, on the Cifar-10 dataset, HLDR and FIMR imparted a small but highly variable improvement in accuracy for all noise powers considered.

Figures 4a - 4d demonstrate how the different models are impacted by the value of ϵ for the AGN-enhanced training procedure described in Section IV. A model trained *only* on AGN-perturbed examples (i.e., $\mathbb{D}_{\mathcal{N}} \setminus \mathbb{D}$) was also evaluated and is distinguished in the legends of 4a-5d as 'AGNT'. AGN training improves accuracy for HLDR in both the AGN and adversarial regimes relative to training on the exclusively adversarial training set, $\mathbb{D}$. AGNT reliably defends against adversarial attacks, but HLDR obtains a significant advantage over AGNT when tested against adversarial noise. In contrast, FIMR and HGD do not experience a significant improvement compared to the first experiment on both adversarial and AGN-perturbed test samples. We also evaluated the output-stability training method described by Zheng et al. [16]. The resulting performances were omitted from these figures because it failed to maintain satisfactory benign test accuracy (specifically by dropping to approximately 10% for every penalty coefficient considered, including those recommended in [16]). Further, stability training was found to be detrimental to adversarial test accuracy (also dropping to approximately 10% in every iteration of CV observed and for every penalty coefficient and test budget considered. Ten such penalty coefficients were sampled uniformly from $[0.0001, 1]$ and found to produce identical results.

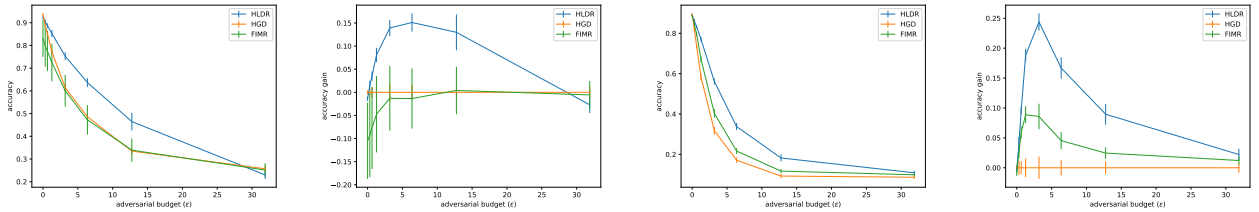
Figures 6a-6d depict the impact on AGN-perturbed test accuracy of training on the combined AGN and adversarially distorted fine-tuning sets, $\mathbb{D}_{\mathcal{N}}$. On both Fashion-Mnist and Cifar-10, this training scheme significantly elevated robustness of HLDR to AGN relative to the undefended model. No apparent improvement between ordinary training and AGN-augmented training was observed for HGD or FIM.

Figures 6a - 6d show the empirical mean perturbation error (i.e., normalized MAE between representations induced in hidden layers of VGG by a benign example and its adversarially perturbed counterpart) for a budget of $\epsilon = \frac{8}{255}$ for the Fashion-Mnist and Cifar-10 datasets under both experiments outlined in section IV. Counter-intuitively, the perturbation errors only differ significantly among the different methodologies on the Fashion-Mnist data after layer six. Their ranking by perturbation error in the second to the last layer of the model does not predict model performance in Figures 2a and 2b. Similarly, in the Cifar-10 experiment, perturbation errors are nearly indistinguishable in the neural network's first eight layers. These observations hold in both experiments described in IV. As in the Fashion-Mnist case, ranking defense technique by perturbation error in latter layers is not predictive of performance on the adversarially perturbed test data shown in Figures 2c and 2d.

Table I shows the time required by each defense technique to iterate over a single sample of fine-tuning. Not shown in the Table I is the observation that the models defended only by AGNT require no additional training time (on a per-sample basis). This stems from the fact that such models are not responsible for the auxiliary computation involved in calculating the regularizers and their derivatives. Notably, the HGD consumes twice as much time as the HLDR and FIMR networks, which differ in training time only by 0.1 milliseconds per sample. This result occurs because HLDR and FIMR do not increase the number of parameters of the defended model. In contrast, HGD increases the load of the forward passes of the optimization routine by more than 11 million parameters. Equivalently, 26028363 parameters must be processed to predict with the HGD defended model.

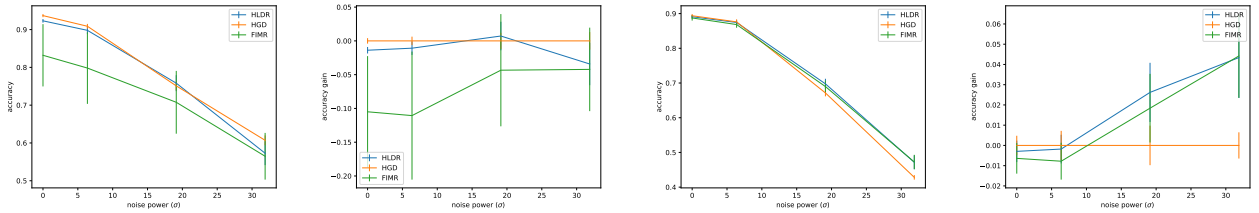
VII. Discussion

Our proposed method has two distinct advantages over the HGD: Aside from significantly better performances in the adversarial test regime, our method only requires as many parameters as the original model, and as a result, completes



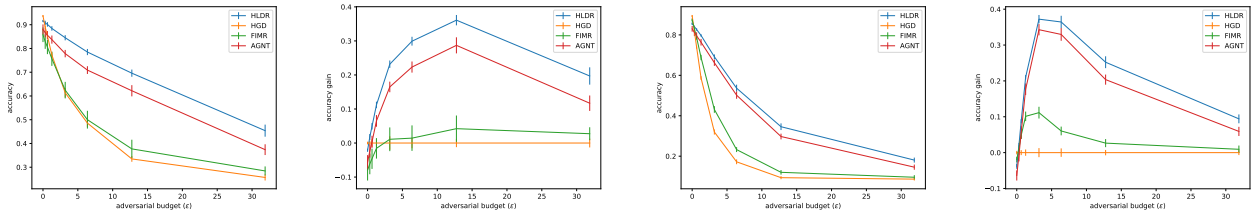
(a) F-mnist Accuracy (b) F-mnist Relative Improvement (c) Cifar-10 Accuracy (d) Cifar-10 Relative Improvement

Fig. 2: (a) Test accuracy vs. adversarial budget for the adversarial defense schemes discussed in this work, evaluated on the Fashion-Mnist dataset (b) Corresponding improvement in test accuracy over the undefended model plotted against adversarial budget. Error bars represent 95% confidence intervals, estimated over ten folds of the CV experiment. (c) Test accuracy vs. adversarial budget for the adversarial defense schemes discussed in this work, evaluated on the Cifar-10 (d) Gain in test accuracy over the undefended model plotted against adversarial budget, evaluated on the Cifar-10 dataset.



(a) F-mnist AGN Accuracy (b) F-mnist AGN Relative Improvement (c) Cifar-10 AGN Accuracy (d) Cifar-10 AGN Relative Improvement

Fig. 3: (a) Test accuracy and (b) the corresponding improvement in test accuracy over the undefended model vs. pixel-wise AGN strength (i.e., standard deviation) measured alongside the results presented in Figures 2a and 2b. (c) and (d) show the analogous results corresponding to the Cifar-10 experiments (i.e., those associated with Figures 2c and 2d.



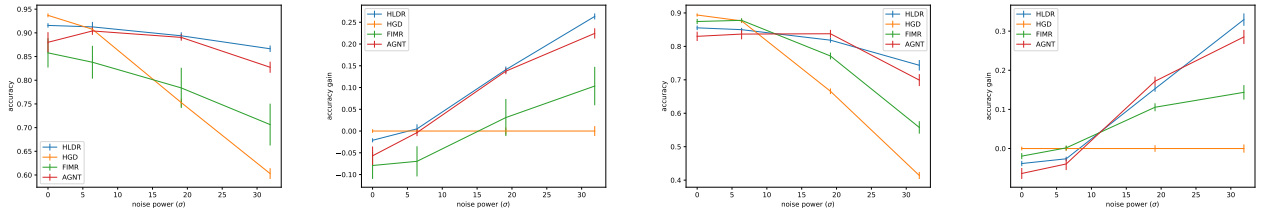
(a) F-mnist Accuracy (b) F-mnist Relative Improvement (c) Cifar-10 Accuracy (d) Cifar-10 Relative Improvement

Fig. 4: (a) Test accuracy vs. adversarial budget for the adversarial defense schemes that were trained on the AGN-enhanced fine tuning subset, $\mathbb{D}_{\mathcal{N}}$, evaluated on the Fashion-Mnist dataset (b) Corresponding improvement in test accuracy over the undefended model plotted against adversarial budget. (c) Test accuracy vs. adversarial budget for the adversarial defense tactics discussed in this work and trained on the AGN-enhanced fine tuning subset, $\mathbb{D}_{\mathcal{N}}$, evaluated on the Cifar-10 (d) Gain in test accuracy over the undefended model plotted against adversarial budget, evaluated on the Cifar-10 dataset.

Training Time ($\frac{\textbf{milliseconds}}{\textbf{sample}}$)		
HLDR	**HGD**	**FIM**
2.1	5.3	2.0

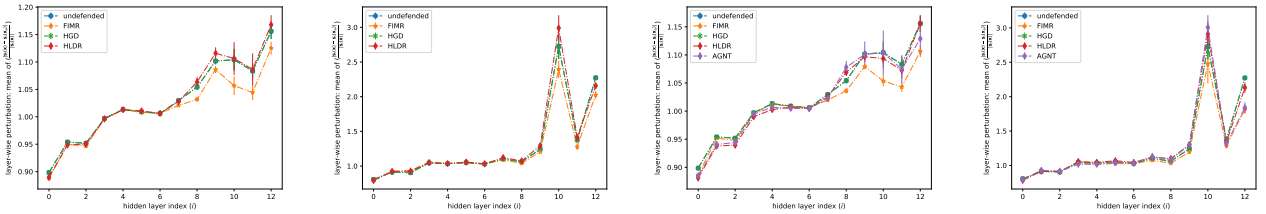
TABLE I: Training Time (measured in seconds per sample) for each of the defense techniques considered in this work. HLDR is nearly as fast as FIM, which (due to its independence from any adversarial information that HLDR must propagate through the network) demonstrates the low cost of the additional forward propagation calculations incurred in order to optimize under the HLDR.

training considerably faster. That FIMR drastically reduces the perturbation induced in hidden layers compared to that of HLDR and HGD is shockingly unintuitive, as one expects that the model with empirically minimal perturbation in the final convolutional layer would be most resistant to adversarial attacks. This discrepancy in the translation of hidden perturbation minimization to adversarial robustness at the output may be explained (in the FIMR and HLDR cases) as resulting from changes in the weights of the final densely connected layers. FIMR reaches a relatively greater robustness to FGSM attacks despite the provision of adversarial examples to HGD and



(a) F-mnist AGN Accuracy (b) F-mnist AGN Relative Improvement (c) Cifar-10 AGN Accuracy (d) Cifar-10 AGN Relative Improvement

Fig. 5: (a) Test accuracy and (b) the corresponding improvement in test accuracy over the undefended model vs. pixel-wise AGN power measured alongside the results presented in Figures 2a and 2b, trained on the AGN-enhanced fine tuning subset, $\mathbb{D}_{\mathcal{N}}$. (c) and (d) show the analogous results corresponding to the Cifar-10 experiments (i.e., those associated with Figures 2c and 2d).



(a) F-mnist Perturbation Error (b) Cifar-10 Perturbation Error (c) F-mnist Perturbation Error (d) Cifar-10 Perturbation Error

Fig. 6: Mean perturbation error - i.e., MAE between hidden representations induced by a benign example and its adversarially perturbed counterpart (with a budget of $\epsilon = \frac{8}{255}$) normalized to the total intensity of the benign representation, $\frac{|\mathbf{s}_i(\mathbf{x}) - \mathbf{s}_i(\mathbf{x}_a)|}{|\mathbf{s}_i(\mathbf{x})|}$ - for hidden convolutional layers of VGG augmented and fine-tuned with each of the various defense techniques considered are shown for the Fashion-Mnist and Cifar-10 datasets in (a) and (b), respectively. (c) and (d) exhibit the same under the combined AGN training regimes.

not to FIMR. This observation demonstrates that the disparity reduction between activity induced in hidden layers by adversarial and benign examples is a more effective means of at once reducing the success rate of adversarial attacks and error amplification (the phenomenon in which this disparity grows with increasing depth in the network in question). That HLDR is able to attain greater improvements in robustness over the baseline demonstrates the value exchanged between the cost of calculating adversarial examples and computing additional forward-passes. We speculate that the gap between adversarial accuracy achieved by HLDR and AGNT is a product of the specific directions in which each training regime expands the boundaries of preimages of decision boundaries embedded in hidden layer activity [32]. Indeed, as AGNT flattens the boundaries of such subsets in the directions that minimize perturbations induced by AGN, in a high dimensional feature space, the sheer number of potential directions exploitable by adversarial perturbations remains large enough that training HLDR attains a tangible (and replicable) improvement in adversarial test accuracy, while matching AGNT's performance on AGN-perturbed test samples.

That we were unable to replicate the robustness conferred by HGD as demonstrated in [15] was unexpected, given its intuitive design and the fact that the structure of its objective function directly inspired the regularizer we introduce. This inconsistency with the results presented in [15] may be attributable to the underlying model. For example, Liao et al.'s investigation studied the defense of a ResNet model while our experiments study VGG [33], [34].

HLDR represents a novel and effective approach to training adversarially robust neural networks; however, our conclusions are limited by the restriction that this work only considers training procedures that use the fine-tuning subsets constructed as described in Section IV. Further work toward understanding how HLDR impacts performance in the face of adversarial, Gaussian, and other forms of distortion (e.g., loss due to compression) involves comparing HLDR directly to other related methods that train on fine-tuning subsets, adversarial attack methods, and objective functions distinct from the formulations used in this work. Of great interest is assessing the impact of replacing the L1-norm in (3) with higher order norms.

ACKNOWLEDGMENT

This work was supported by grants from the Department of Energy #DE-NA0003946, and National Science Foundation (NSF) CAREER #1943552. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Haris Iqbal's open source software, PlotNeuralNet, was used in combination with Inkscape to illustrate Figure 1. Source code and pre-trained models can be accessed at:
https://github.com/dmschwar/robust-ldr

REFERENCES

[1] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571, IEEE, 2016.

[2] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural computation*, vol. 29, no. 9, pp. 2352–2449, 2017.

[3] W. Alakwaa, M. Nassef, and A. Badr, "Lung cancer detection and classification with 3d convolutional neural network (3d-cnn)," *Lung Cancer*, vol. 8, no. 8, p. 409, 2017.

[4] J. Guo, Z. Liang, E. Scribner, G. Ditzler, N. Bouaynaya, and H. Fathallah-Shaykh, "Nonlinear brain tumor model estimation with long short-term memory neural networks," in *IEEE/INNS International Joint Conference on Neural Networks*, 2018.

[5] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *2013 IEEE international conference on acoustics, speech and signal processing*, pp. 7398–7402, IEEE, 2013.

[6] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, *et al.*, "Mastering the game of go without human knowledge," *nature*, vol. 550, no. 7676, pp. 354–359, 2017.

[7] K. Sadeghi, A. Banerjee, and S. K. Gupta, "A system-driven taxonomy of attacks and defenses in adversarial machine learning," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2020.

[8] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," *arXiv preprint arXiv:1611.01236*, 2016.

[9] J. Gilmer, N. Ford, N. Carlini, and E. Cubuk, "Adversarial examples are a natural consequence of test error in noise," in *International Conference on Machine Learning*, pp. 2280–2289, 2019.

[10] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial examples are not bugs, they are features," in *Advances in Neural Information Processing Systems*, pp. 125–136, 2019.

[11] A. Rozsa, M. Günther, and T. E. Boult, "Are accuracy and robustness correlated," in *2016 15th IEEE international conference on machine learning and applications (ICMLA)*, pp. 227–232, IEEE, 2016.

[12] F. Tramèr, J. Behrmann, N. Carlini, N. Papernot, and J.-H. Jacobsen, "Fundamental tradeoffs between invariance and sensitivity to adversarial perturbations," *arXiv preprint arXiv:2002.04599*, 2020.

[13] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli, "Evasion attacks against machine learning at test time," in *Joint European conference on machine learning and knowledge discovery in databases*, pp. 387–402, Springer, 2013.

[14] Y. Liu, Y. Xie, and A. Srivastava, "Neural trojans," in *2017 IEEE International Conference on Computer Design (ICCD)*, pp. 45–48, IEEE, 2017.

[15] F. Liao, M. Liang, Y. Dong, T. Pang, X. Hu, and J. Zhu, "Defense against adversarial attacks using high-level representation guided denoiser," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1778–1787, 2018.

[16] S. Zheng, Y. Song, T. Leung, and I. Goodfellow, "Improving the robustness of deep neural networks via stability training," in *Proceedings of the ieee conference on computer vision and pattern recognition*, pp. 4480–4488, 2016.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[18] K. Meshkini, J. Platos, and H. Ghassemain, "An analysis of convolutional neural network for fashion images classification (fashion-mnist)," in *International Conference on Intelligent Information Technologies for Industry*, pp. 85–95, Springer, 2019.

[19] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do cifar-10 classifiers generalize to cifar-10?," *arXiv preprint arXiv:1806.00451*, 2018.

[20] M. Cheng, Q. Lei, P.-Y. Chen, I. Dhillon, and C.-J. Hsieh, "Cat: Customized adversarial training for improved robustness," *arXiv preprint arXiv:2002.06789*, 2020.

[21] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[22] F. Carrara, R. Becarelli, R. Caldelli, F. Falchi, and G. Amato, "Adversarial examples detection in features distance spaces," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 0–0, 2018.

[23] C. Shen, Y. Peng, G. Zhang, and J. Fan, "Defending against adversarial attacks by suppressing the largest eigenvalue of fisher information matrix," *arXiv preprint arXiv:1909.06137*, 2019.

[24] L. L. Scharf, *Statistical Signal Processing*. Addison-Wesley, 1991.

[25] V. Zantedeschi, M.-I. Nicolae, and A. Rawat, "Efficient defenses against adversarial attacks," in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pp. 39–49, 2017.

[26] D. Riach, "Determinism in deep learning," *ScaledML*, 2019.

[27] S. Jean-Paul, T. Elseify, I. Obeid, and J. Picone, "Issues in the reproducibility of deep learning results," in *2019 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pp. 1–4, IEEE, 2019.

[28] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[29] A. Krizhevsky, G. Hinton, *et al.*, "Learning multiple layers of features from tiny images," 2009.

[30] T. Dozat, "Incorporating nesterov momentum into adam," *International Conference on Learning Representations (ICLR)*, 2016.

[31] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, pp. 448–456, PMLR, 2015.

[32] A. Fawzi, S.-M. Moosavi-Dezfooli, P. Frossard, and S. Soatto, "Empirical study of the topology and geometry of deep networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3762–3770, 2018.

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.