Towards Efficient and Effective Adversarial Training

Anonymous Author(s)
Affiliation
Address
email

Abstract

The vulnerability of Deep Neural Networks to adversarial perturbations has spurred immense interest towards the development of models that are robust to such attacks. However, present state-of-the-art adversarial defenses involve the use of 10-step adversaries, which renders them computationally infeasible for application to large-scale datasets. While recent single-step adversarial training methods show promising direction, their robustness is not on par with multi-step training methods. In this work, we bridge this performance gap by introducing a novel Nuclear-Norm regularizer on network predictions to enforce function smoothing in the vicinity of data samples, and boost the accuracy further by using Weight Averaging for both attack generation and training. While prior works consider each data sample independently, the proposed Nuclear-Norm Adversarial Training (NuAT) imposes a rank minimization constraint on the oscillation of function values across a training minibatch, thereby enhancing optimization using the joint batch-statistics of adversarial samples. We achieve state-of-the-art results amongst existing efficient training methods and demonstrate results which are comparable to that of multi-step adversarial training methods such as TRADES and PGD-AT, at a significantly lower computational cost.

1 Introduction

The past decade has witnessed the rise of Deep Learning based systems and solutions in many applications related to Computer Vision, Natural Language Processing and Speech Processing. Despite the remarkable success of Deep Networks, they are known to be susceptible to crafted imperceptible noise called Adversarial Attacks [26], which can have disastrous implications when deployed in critical applications such as self-driving cars, medical diagnosis and surveillance systems. The robustness of Deep Networks to Adversarial Attacks has been of prime importance, since this relates to improving their worst-case performance [5].

Early attempts of improving robustness to adversarial attacks could be broadly classified into input pre-processing based defenses [3, 31, 24, 12] and adversarial training methods [10, 19]. While the input pre-processing based defenses were computationally cheap, they primarily involved masking of gradients to prevent the generation of strong attacks. Such methods were later shown to be ineffective against attacks constructed using expectation over the randomized components or smooth approximations of non-differentiable components [2]. Amongst the latter methods, the most successful early defense which stood the test of time was the Projected Gradient Descent (PGD) adversarial training method proposed by Madry et al. [19]. This involved minimization of cross-entropy loss on the worst-case perturbations generated using multiple iterations of constrained optimization, leading to a significantly higher computational cost when compared to the standard training.

FGSM (Fast Gradient Sign Method) based adversarial training [10] alleviates the computational cost by utilizing single-step adversarial samples for training. However, this method is contingent on the local linearity assumption of the loss surface, which is often compromised during the process of

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

FGSM training, leading to generation of weaker adversaries as training progresses. This leads to the generation of models which exhibit a false sense of robustness towards single-step attacks and are susceptible to stronger multi-step attacks. Although there have been several recent works [28, 29, 25] to circumvent gradient masking and improve the worst-case accuracy of single-step adversarial training methods, the robustness thus achieved has been suboptimal compared to multi-step defenses.

In this work, we propose a novel Nuclear-Norm regularizer to impose local smoothness in the vicinity of data samples, thereby improving the effectiveness of single-step adversarial training, and further boost the accuracy by using weight averaging for both attack generation and training. The proposed Nuclear-Norm Adversarial Training (NuAT) utilizes batch statistics to limit the oscillation of function values, thereby preventing the over-smoothing of loss surface uniformly in all directions. The contributions of our work can be summarized as follows:

- We propose single-step Nuclear Norm Adversarial Training (NuAT), and a two step variant of the same (NuAT2), coupled with a novel cyclic-step learning rate schedule, to achieve state-of-the-art results amongst the *efficient* training methods, and results comparable to the *effective* multi-step training methods.
- We propose the use of exponential (model) weight averaging coupled with cyclic learning rate schedule (NuAT-WA) to boost the accuracy of NuAT further.
- We propose NuAT2-WA, which utilizes stable gradients from the weight-averaged model to obtain a better initialization in the first attack step, leading to improved robustness for the same training budget.
- We propose to brigde the gap between single-step and two-step training by introducing a hybrid approach that switches adaptively from single-step to two-step training to stabilize single-step training, and boost its performance further.

The organization of this paper is as follows: The preliminaries on notation and threat model are laid out in Section-2, followed by a brief note on related works in Section-3 and a discussion on challenges in single-step adversarial training in Section-4. We present a detailed description of our proposed method in Section-5, followed by experimental results to support our claims in Section-6. We conclude the paper with our analysis and future directions in Section-7.

2 Preliminaries

In this paper, we denote x to be a d-dimensional image from an N-class dataset $\mathcal{D}$. Further, we denote its corresponding ground-truth label as a one-hot vector y. Let f_θ represent a Deep Neural Network with parameters θ, that maps an input image x to its pre-softmax output $f_\theta(x)$. The cross-entropy loss corresponding to the network prediction on a sample (x, y) is denoted as $\ell_{CE}(f_\theta(x), y)$.

For a minibatch $B = \{(x_i, y_i)\}_{i=1}^{M}$, we denote X as the image matrix whose i^{th} row consists of image x_i, and denote Y as the corresponding ground-truth array. Thus, X is a matrix of size $(M \times d)$, and $f_\theta(X)$ represents a prediction matrix of size $(M \times N)$. Let $\ell_{CE}(f_\theta(X), Y)$ denote the sum of cross-entropy losses over all data samples in the minibatch B. Further, for a matrix A, let $||A||_*$ denote the Nuclear Norm, that is the sum of the singular values of A.

Adversarial Threat Model: In this paper, we seek to study the robustness of Deep Networks in the standard setting of ℓ_∞ constrained adversaries. Thus under an ε-constraint, given a clean image x, an adversarially perturbed counterpart $\widetilde{x}$ can differ by at most ε at any given pixel location. Further, a network f_θ is said to be ε-robust on a clean sample x with label y, if $f_\theta(\widetilde{x}) = y$, for all perturbations $\widetilde{x}$ such that $||x - \widetilde{x}||_\infty \leq \varepsilon$.

3 Related Works

Multi-step Adversarial Defenses: Amongst the most effective defenses known till date is that of PGD adversarial training as introduced by Madry et al. [19], wherein multi-step optimization is performed following the addition of random noise to generate strong adversaries resistant to gradient masking. Further, Zhang et al. [32] proposed TRADES, a multi-step adversarial method that balances the trade-off between standard accuracy and adversarial robustness. By training on adversaries generated by maximising the Kullback-Leibler (KL) Divergence, smooth models could be reliably

obtained, and was seen to achieve state-of-the-art robust performance on several benchmarks. Rice et al. [22] found that early stopping could be used to alleviate the phenomenon of robust-overfitting, wherein the robust performance of deep networks degrades as training is continued beyond a point. Strikingly, it was observed that early-stopping combined with PGD adversarial training could be used to produce models that were significantly more robust, and can even out-perform TRADES based training by utilizing larger model capacities. In more recent work, Pang et al. [20] discuss several implementation details with regard to the optimal choice of hyperparameters, and further improve upon the hitherto TRADES defense.

Defenses using Weight Regularization: Wu et al. [30] explore the use of Adversarial Weight Perturbations (AWP) in order to explicitly enforce a flatter weight loss landscape. By generating perturbations in both input and weight space during the training regime, the combined adversaries so found maximise the loss with respect to both domains. AWP is seen to significantly improve robust performance, albeit with an increased computational cost and a copious set of additional hyperparameters. Prior works such as Izmailov et al. [13] have shown that Stochastic Weight Averaging (SWA) can also be used to train networks with flatter and wider minima with improved generalisation in the context of standard training and test performance, as compared to solutions obtained using traditional Stochastic Gradient Descent (SGD). In a very recent work, Gowal et al. [11] investigate the effect of factors such as choice of activation function, additional unlabelled data, as well as the application of exponential weight averaging to robust multi-step defenses to boost adversarial performance. In contrast, in this work we propose to additionally incorporate adversaries generated from the Stochastic Weight Averaged model in order to generate diverse and effective perturbations. Thus, by additionally utilising supervision from the weight-averaged model during 2-step adversarial training (NuAT2-WA), the efficacy of the defense is significantly enhanced.

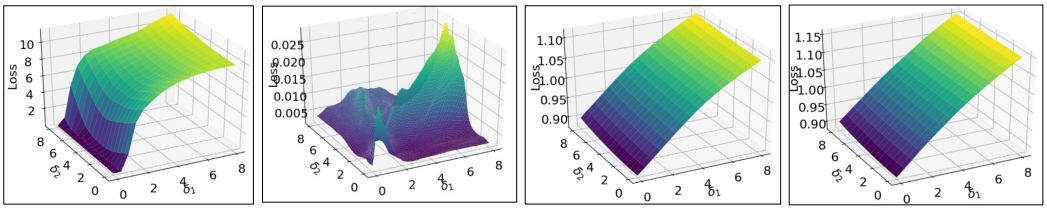
Efficient Adversarial Training: While multi-step adversarial methods continue to achieve the best robust performance, their computational complexity poses a severe impediment to their deployment in real-world applications that utilize large-scale datasets. Several works have been proposed to efficiently train robust models, either by reusing gradient computations [23], or by utilising single-step adversarial training. We expound upon the latter method in greater detail in the following section.

Nuclear Norm Regularization: Optimization objectives utilizing the Nuclear norm regularizer have seen extensive use in several computer vision, compressed sensing and machine learning applications such as in image denoising [14], low-rank matrix completion [4] and multi-task learning [21]. The Nuclear norm provides a tight convex relaxation for the rank of a matrix, and can thus be successfully used as a surrogate objective to approximately solve intractable optimization problems. In this work, we seek to utilize Nuclear norm regularization to enforce consistent network predictions within the vicinity of clean data samples through adversarial training, and is discussed in detail in Section-5.

4 Analysis of Single-Step Adversarial Training

FGSM Adversarial Training: The efficacy of single-step approximations for generating adversarial attacks was first demonstrated by Goodfellow et al. [10]. The single-step FGSM attack was found to be sufficient to compromise the accuracy of models trained using empirical risk minimization. These single-step adversaries were further used for adversarial training (FGSM-AT), where the original dataset is augmented with some proportion of FGSM adversarial samples. However, it was later found that the resulting models were not robust to multi-step attacks due to the gradient masking effect [16]. The FGSM-AT loss formulation, which combines cross-entropy loss on clean and single-step adversarial samples, has two possible solutions. The first is the desired solution, where the loss surface of the trained network continues to remain locally linear as training progresses, resulting in the generation of strong single-step adversaries, thereby facilitating the training of robust models. The second is a degenerate solution, where the model develops a convoluted loss surface, which undermines the linearity assumption required for the generation of effective single-step adversaries. Since the adversaries used for training are not effective, the adversarially trained model is not robust to multi-step attacks. This indeed results in a better minimizer, since training on weak adversaries would lead to lower cross-entropy loss on both clean and adversarial samples.

In practice, as training progresses, the loss surface starts deviating increasingly from the first-order assumption, thereby reducing the efficacy of FGSM samples and leading the training process to converge to the degenerate solution. Fig.1(b) shows the loss surface of an FGSM trained model, where



(a) Normal Training (b) FGSM-AT (c) PGD-AT (d) NuAT(Ours)

Figure 1: Loss surface plots showing the perturbations of the form, $x^* = x + \delta_1 g + \delta_2 g^{\perp}$, where g is the sign of the gradient direction of the cross-entropy loss and $g^{\perp}$ is an orthogonal direction.

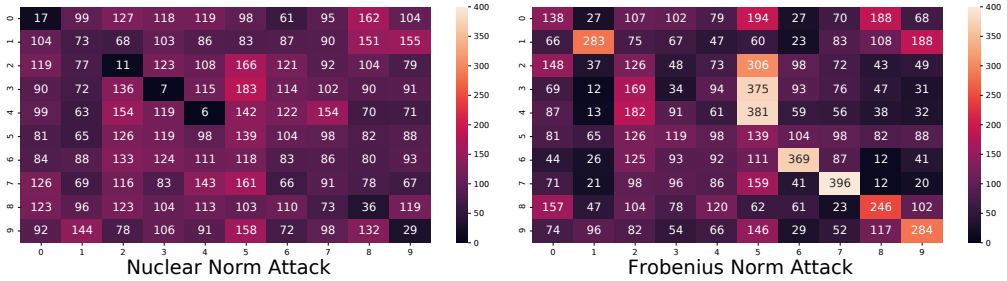


Figure 2: **Confusion Matrices** for predictions against adversarial attacks generated by maximizing the Nuclear norm and Frobenius norm of a matrix respectively. These are obtained for a normally trained model with ResNet-18 architecture on CIFAR-10 dataset.

the gradients at the vicinity of the sample are masked, thereby preventing single-step methods from reaching a true maximum in the ε-ball. It can be observed that traversal along the gradient direction (δ_1) does not result in increased loss, preventing the generation of strong single-step adversaries.

R-FGSM Adversarial Training: R-FGSM adversarial training [27] attempts to escape the gradient masking effect by computing gradients at a random point in the vicinity of the data sample. This relies on the assumption that gradients would remain constant in the local neighbourhood of data samples when the loss surface is smooth. However, this assumption can fail in case of gradient masking, leading to the generation of weak adversaries, which is termed as catastrophic overfitting in subsequent works [29]. Although Wong et al. [29] demonstrated successful R-FGSM training by adding larger magnitudes of noise before attack (FBF), and using shorter training schedules, the models trained were not competent with multi-step adversarial training methods. Also, this method does not improve robustness on larger models like WideResNet, as shown in subsequent works [25].

Regularizers to Mitigate Gradient Masking: One of the effective methods of mitigating gradient masking has been the use of regularizers to improve local smoothness of the loss surface [28, 25]. Sriramanan et al. [25] show that imposing an ℓ_2 regularizer on the softmax values of clean samples and their corresponding single-step adversaries can result in significant gains in adversarial robustness when compared to R-FGSM training [29]. While this has proved to be the most effective strategy for single-step adversarial training, the results obtained are still not on par with multi-step adversarial training methods. The ℓ_2 regularizer introduced in GAT [25] attempts to uniformly limit the oscillation of the network outputs along all dimensions. This imposes an overly severe constraint on the outputs of the network, thereby limiting the network from learning the true underlying function. This results in a larger robustness-accuracy tradeoff when compared to multi-step adversarial training. It is to be noted that function smoothness is a result of PGD adversarial training as well (Fig.1(c)). However, the absence of an explicit regularizer allows more degrees of freedom for the function learned.

In this work, we introduce a novel Nuclear-Norm based regularizer to utilize batch-statistics for achieving function smoothing only in the required dimensions, and to the required extent. This prevents the over-smoothing of the loss function, thereby resulting in significant improvement in the robustness achieved. We discuss the proposed Nuclear-Norm Adversarial Training (NuAT) algorithm and analyse properties of models trained using this regularizer in the following section.

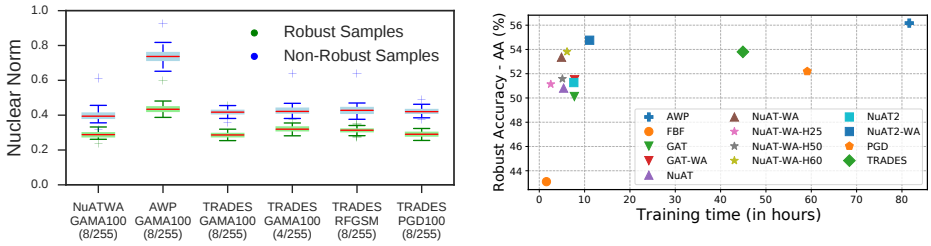


Figure 3: (a) Nuclear Norm for Robust and Non-Robust samples for multiple defense methods (b) **Computational Complexity:** Plot of Robust Accuracy on AutoAttack (AA-v2) versus training time (in hours) for multiple defense methods.

5 Proposed Method

5.1 Nuclear-Norm based Attack

We consider X to be the matrix composed of vectorized pixel values (arranged row-wise) of each image in a given minibatch B of size M. We first add independently sampled Bernoulli noise of magnitude α, pixel-wise to each image. We consider the matrix Δ to be composed of these random perturbations added to each image. The addition of noise ensures that the model does not rely heavily on gradients at the data samples, thereby preventing the occurrence of gradient masking. The following loss function is then maximized for the generation of single-step adversaries:

$$\widetilde{L} = \ell_{CE}(f_\theta(X+\Delta), Y) + \lambda \cdot ||f_\theta(X+\Delta) - f_\theta(X)||_* \tag{1}$$

The matrix Y is composed of the ground-truth one-hot vectors for the corresponding images in X. The first term in Eq.1 represents the cross-entropy loss on the clean samples perturbed with Bernoulli noise, and the second term is the Nuclear Norm of difference in pre-softmax values $f_\theta(.)$ of the clean images and the corresponding perturbed images. The loss in Eq.1 is maximized to find the optimal perturbation Δ within the defined ℓ_∞ constraint set.

Maximization of nuclear norm leads to an increase in singular values of the matrix $f_\theta(X+\Delta)-f_\theta(X)$, resulting in more diverse attacks. The generated adversarial attacks are encouraged to go to a diverse set of classes, as opposed to the most confused class which is a common target in standard cross-entropy based attacks, or in attacks which maximize Frobenius norm. We compare the confusion matrices for predictions against adversarial attacks generated by maximizing the Nuclear norm and Frobenius norm of a matrix respectively in Fig.2. These are obtained for a normally trained model with ResNet-18 architecture on CIFAR-10 dataset. For a better understanding, we use the same class images in a single batch (with batch size $= 100$) for generating the attack. It can be observed that the Nuclear Norm based attack results in a better diversification of the target classes after attack when compared to the Frobenius Norm attack. We also quantify this difference by computing the entropy of the average softmax vector for each class, and further computing an average entropy score across all classes. We obtain an average entropy score of 0.87 and 0.97 for the Frobenius Norm attack and Nuclear Norm attack respectively, clearly indicating better diversification in the latter attack. It can be observed from the diagonal entries of the confusion matrix that Nuclear Norm maximization also results in a stronger attack when compared to the Frobenius Norm attack.

While maximization of the cross-entropy term in Eq.1 aids in finding a strong attack, the Nuclear Norm regularizer helps mitigate gradient masking, while also encouraging better diversity in the generated attacks. The Nuclear-Norm attack is stronger against FGSM-AT model, since it exhibits gradient masking, whereas R-FGSM based attack is stronger for models where the loss surface is smooth (TRADES) (Details in the Supplementary). Therefore, both R-FGSM (CE-only) attack and Nuclear-Norm attack play an important role during single-step adversarial training. While R-FGSM generates a strong attack for robust training, it is susceptible to gradient-masking effects. On the other hand, addition of the Nuclear-Norm regularizer improves the local smoothness of the loss surface, thereby resulting in a better gradient direction for the attack. We therefore set the coefficient of the Nuclear-Norm term to 0 in alternate minibatches, in order to utilize R-FGSM attack and Nuclear-Norm attack alternately for training. This also diversifies the attacks used for adversarial training, as noted by Sriramanan et al. [25].

Algorithm 1 Nuclear Norm Adversarial Training

Input: Network f_θ with parameters θ, Training Data $\mathcal{D}$ with input images of dimension d, Minibatch Size M, Attack Size ε, Initial Noise Magnitude α, Epochs E, Learning Rate η
for $epoch = 1$ **to** E **do**
  for minibatch $B = \{(x_i, y_i)\}_{i=1}^{M} \subset \mathcal{D}$ **do**
    $X = \begin{bmatrix} \dots & x_1 & \dots \\ \dots & \vdots & \dots \\ \dots & x_M & \dots \end{bmatrix}, \quad \Delta = \begin{bmatrix} \dots & \delta_1 & \dots \\ \dots & \vdots & \dots \\ \dots & \delta_M & \dots \end{bmatrix}, \quad \delta_i \sim Bern^d(-\alpha, \alpha), \quad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_M \end{bmatrix}$
    $\widetilde{L} = \ell_{CE}\left(f_\theta(X + \Delta), Y\right) + \lambda \cdot ||f_\theta(X + \Delta) - f_\theta(X)||_*$
    $\Delta = \Delta + \varepsilon \cdot \text{sign}\left(\nabla_\Delta \widetilde{L}\right)$
    $\Delta = Clamp\left(\Delta, -\varepsilon, \varepsilon\right), \quad \widetilde{X} = Clamp\left(X + \Delta, 0, 1\right)$
    $L = \ell_{CE}(f_\theta(X), Y) + \lambda \cdot ||f_\theta(\widetilde{X}) - f_\theta(X)||_*$
    $\theta = \theta - \frac{1}{M} \cdot \eta \cdot \nabla_\theta L$
  end for
end for

5.2 Nuclear-Norm Adversarial Training

The algorithm of the proposed Nuclear-Norm Adversarial Training is presented in Algorithm-1. The following loss function is minimized for Nuclear-Norm Adversarial Training:

$$L = \ell_{CE}(f_\theta(X), Y) + \lambda \cdot ||f_\theta(\widetilde{X}) - f_\theta(X)||_* \tag{2}$$

Here $\widetilde{X}$ is a matrix consisting of adversarial perturbations corresponding to the images in X. As discussed, these perturbations are generated using R-FGSM attack and Nuclear-Norm based attack in alternate iterations of training. The first term in Eq.2 corresponds to the cross-entropy loss on clean samples, and the second term corresponds to Nuclear-Norm of difference in pre-softmax values of clean images X and their corresponding single-step adversaries $\widetilde{X}$. We utilize the same loss function for both adversary generation and training as used by Madry et al. [19]. The proposed regularizer in Eq.1 and 2 is weighted by a factor λ, which controls the accuracy-robustness trade-off. We use the same weight λ for both adversary generation and training, which is linearly increased over training epochs. We show the sensitivity of the training to the parameter λ in the Supplementary material. Further, we utilise a novel learning rate schedule that incorporates the cyclic schedule in early epochs of training, and transitions to the use of a step schedule towards the end of the training regime (details in the Supplementary Section).

While the Nuclear Norm attack encourages the generation of perturbations which maximize the singular values of the matrix $f_\theta(\widetilde{X}) - f_\theta(X)$, the NuAT defense attempts to minimize the same. A minimizer of the Nuclear Norm is one which additionally minimizes the rank of the matrix, thereby encouraging associations between the columns (or classes) of the matrix (data). This allows the samples within a batch to aid each other and utilize the relationship between the oscillation of the function values across different dimensions (classes) to obtain a better robustness-accuracy trade-off.The Nuclear Norm of difference in pre-softmax values of X and $\widetilde{X}$ is significantly higher for a normally trained model when compared a robust model (Plot in Supplementary). As shown in Fig.3(a), across multiple robust models, lower Nuclear-Norm values clearly correspond to robust samples and the higher values correspond to non-robust ones. This reiterates our hypothesis that explicit minimization of the Nuclear Norm regularizer can improve the adversarial robustness of models.

Loss Surface Plots: We visualize the loss surface plots for models trained using different training methods in Fig.1. We plot the cross-entropy loss by varying the perturbation along two axes: one along the sign of the gradient, and another along a random direction orthogonal to the gradient. As expected, we observe that FGSM adversarial training produces extreme gradient masking in Fig.1(b), resulting in a convoluted loss landscape. On the other hand, the loss surface of the proposed single-step defense Nuclear Norm Adversarial Training (NuAT) is seen to be smooth in Fig.1(d),

Table 1: **CIFAR-10 White-box evaluation:** Accuracy (%) of various defenses (rows) against different attacks for the CIFAR-10 dataset, under an ℓ_∞ constraint of $\varepsilon = 8/255$. The defenses presented are as follows: FGSM-AT [10], RFGSM-AT [27], ATF [23], FBF [29], R-MGM [28], GAT [25], PGD-AT [19, 22], TRADES [32, 20] and AWP [30].

(a) **ResNet-18 architecture**

Method	**# AT steps**	**Clean Acc**	**PGD (n-steps) 20**	**PGD (n-steps) 500**	**GAMA 100**	**AA** (v1)
Normal	0	92.30	0.00	0.00	0.00	0.00
FGSM-AT	1	**92.89**	0.00	0.00	0.00	0.00
RFGSM-AT	1	89.24	35.02	34.17	33.87	33.16
ATF	1	71.77	43.53	43.52	40.34	40.22
FBF	1	82.83	46.41	46.03	43.85	43.12
R-MGM	1	82.29	46.23	45.79	44.06	43.72
GAT	1	80.49	53.13	53.08	47.76	47.30
GAT-WA	1	79.47	**54.40**	**54.37**	49.00	48.28
NuAT (**Ours**)	1	81.01	53.30	52.97	49.46	49.24
NuAT-WA (**Ours**)	1	82.21	54.14	53.95	**50.97**	**50.75**
PGD-AT	10	81.12	53.08	52.89	49.08	48.75
TRADES	10	81.47	52.73	52.61	49.22	49.06
TRADES-WA	10	80.19	52.98	52.88	49.49	49.39
AWP	11	**81.99**	**55.60**	**55.52**	**51.65**	**51.45**

(b) **WideResNet-34-10 architecture**

Method	**AT-steps (epochs)**	**Clean Acc**	**PGD 100**	**GAMA 100**	**AA (v2)**
FBF	1 (30)	82.05	45.57	43.13	43.10
GAT	1 (85)	85.17	55.12	50.76	50.12
GAT-WA	1 (85)	84.61	57.28	52.19	51.50
Variants of NuAT (**Ours**)					
NuAT	1 (55)	85.30	53.82	51.34	50.81
NuAT-H	1 (50_{+5})	84.58	54.89	51.93	51.58
NuAT-WA	1 (50)	85.29	56.21	53.73	53.36
NuAT-WA-H	1 (25_{+2})	81.98	54.82	51.41	51.14
NuAT-WA-H	1 (60_{+6})	84.93	57.51	54.28	53.81
NuAT2	2 (55)	84.76	54.50	51.99	51.27
NuAT2-WA	2 (80)	**86.32**	**57.74**	**55.08**	54.76
TRADES	10 (110)	85.48	56.35	53.88	53.80
PGD	10 (200)	**86.07**	55.74	52.70	52.19
AWP	11 (200)	85.36	**59.13**	**56.35**	**56.17**

indicating the absence of gradient masking. We further note that the smoothness of loss surface of NuAT is comparable to that of PGD-AT, which is a multi-step adversarial training method.

5.3 Model Weight Averaging

Stochastic Weight Averaging [13, 6] is known to produce flatter and wider optima when compared to Stochastic Gradient Descent (SGD) leading to better generalization. Large capacity networks such as WideResNet are known to exhibit the phenomenon of robust overfitting [22] during adversarial training, and are observed to possess loss landscapes with sharp local minima. We thus utilise exponential weight averaging in order to produce smooth, flatter optima that yield loss landscapes that closely adhere to first-order linear approximations locally, thereby enhancing the efficacy of single-step adversarial training. We integrate this approach with the proposed NuAT single-step defense, and further propose a novel two-step defense NuAT2-WA, which utilizes the weight-averaged model for attack generation. We present further details in Supplementary Material.

6 Experiments and Analysis

Overview of Datasets and Evaluations: We run extensive evaluation on the following three benchmark datasets: MNIST [17], CIFAR-10 [15] and a 100 class subset of ImageNet [9, 25]. For each dataset, we maintain a train-validation split that is balanced equally across all classes. We present details on the datasets, train-validation splits and model architectures in the Supplementary Section. In this Section, we present evaluations on a wide range of adversaries in a white-box setting. We consider an ℓ_∞ adversarial constraint of $\varepsilon = 8/255$ for CIFAR-10 and ImageNet-100, and of $\varepsilon = 0.3$ for MNIST, as is standard. We first present results using untargeted attacks, over a broad range of adversaries such as PGD [19] and GAMA-PGD [25]. We additionally present results on AutoAttack [8], consisting of an ensemble of the APGD-CE, APGD-DLR, FAB [7] and Square [1] attacks. We remark that the GAMA-PGD and AA attacks comprise the strongest known attacks till date, and provide a reliable estimate of the true robustness of Deep Networks.

6.1 White-Box Attacks

CIFAR-10 (ResNet-18): In Table-1a, we present an evaluation of the proposed method on the CIFAR-10 dataset, and compare it with several existing defenses using the ResNet-18 architecture. We observe that the proposed method NuAT performs significantly better than existing single-step training methods such as FGSM-AT [10], RFGSM-AT [27], ATF [23], FBF [29], R-MGM [28] and GAT [25]. Further, we note that despite using only single-step adversaries for training, NuAT performs better than multi-step defenses such as PGD adversarial training (PGD-AT) [19, 22] and TRADES [32, 20]. We observe that by incorporating exponential weight-averaging [13] with the proposed single-step defense, NuAT-WA is seen to achieve enhanced robust performance, along with

Table 2: **Consolidated results against white-box:** Prediction accuracy (%) of different models against PGD 500-step, GAMA-PGD 100-step, AutoAttack-v1 (AA) attacks across CIFAR-10, ImageNet-100 and MNIST datasets. The defenses presented are: RFGSM-AT [27], FBF [29], R-MGM [28], GAT [25], PGD-AT [19, 22] and TRADES [32, 20].

	CIFAR-10				**ImageNet-100**				**MNIST**			
	Clean Acc	**PGD 500**	**GAMA 100**	**AA (v1)**	**Clean Acc**	**PGD 500**	**GAMA 100**	**AA (v1)**	**Clean Acc**	**PGD 500**	**GAMA 100**	**AA (v1)**
Normal	**92.30**	0.00	0.00	0.00	**81.44**	0.00	0.00	0.00	99.20	0.00	0.00	0.00
RFGSM-AT	89.24	34.17	33.87	33.16	78.46	13.88	13.38	12.96	**99.37**	85.32	83.64	82.28
FBF	82.83	46.03	43.85	42.37	57.32	27.22	21.78	20.66	99.30	91.37	87.27	79.02
R-MGM	82.29	45.79	44.06	43.72	64.84	31.68	27.46	27.68	99.04	90.56	88.13	86.21
GAT	80.49	53.08	47.76	47.30	67.98	37.46	29.30	28.92	**99.37**	94.44	92.96	90.62
NuAT (**Ours**)	81.01	52.97	49.46	49.24	69.00	37.60	32.38	31.96	**99.37**	96.24	94.65	**93.11**
NuAT-WA (**Ours**)	82.21	**53.95**	**50.97**	**50.75**	68.40	**38.68**	**33.22**	**33.16**	99.36	**96.30**	**94.70**	93.10
TRADES	81.47	52.61	49.22	49.06	62.88	37.24	31.44	31.66	99.32	93.40	92.74	92.19
PGD-AT	81.12	52.89	49.08	48.75	68.62	36.56	32.24	32.98	99.27	93.98	92.80	91.81

improved clean accuracy. Further on both these metrics, NuAT-WA significantly outperforms other methods with weight-averaging incorporated such as GAT-WA and TRADES-WA. We also note that NuAT-WA achieves robust accuracy comparable to that of AWP [30], while requiring only a fraction of the budget to generate adversaries during training. Over five runs with random seeds, we record a standard deviation of 0.11 and 0.16 respectively for the clean and 100-step GAMA-PGD accuracies.

CIFAR-10 (WRN-34-10): In Table-1b, we compare the proposed defense with existing approaches on the WideResNet-34-10 architecture. We note that that NuAT significantly outperforms the FBF model, since the latter suffers from catastrophic overfitting on larger model capacities. We also remark that NuAT achieves better robustness, and is able to strike a slightly more favourable robustness-accuracy tradeoff as compared to GAT, while requiring far fewer epochs for training as well. Next, we introduce 2-step Nuclear Norm Adversarial Training (NuAT2), wherein the first step is generated in a manner identical to that of NuAT, and followed by the maximisation of cross-entropy loss in the second. This procedure generates diverse adversaries and is seen to significantly improve robustness.

Weight Averaging in Wide Networks: Indeed, in Table-1b, we observe a remarkable improvement with NuAT-WA, while requiring little computational overhead. The NuAT-WA model is seen to significantly outperform 10-step PGD adversarial training, and achieve robustness comparable to that of TRADES while requiring significantly less computational time for training. We also present results by incorporating supervision from the exponentially weight-averaged model with 2-step Nuclear Norm Adversarial Training (NuAT2-WA) to enhance attack strength and diversity. Here, we first generate a single-step attack by maximising the proposed loss (Eq.1) on the exponential weight-averaged model, and then utilise a second step that maximizes cross-entropy loss on the original base-model to obtain strong adversaries. NuAT2-WA amply surpasses the TRADES defense both in terms of robust performance, as well as clean accuracy, and achieves robustness comparable to that of 11-step AWP training. We discuss the training methodology in greater detail and present similar detailed results on ImageNet-100 and MNIST in the Supplementary Material.

Consolidated results across three datasets: In Table-2, we present consolidated results across all three datasets, CIFAR-10, ImageNet-100 and MNIST using the set of the strongest known white-box attacks. We primarily utilise the baseline models as introduced by Sriramanan et al. [25] to perform a comparative analysis for the latter two datasets. Though we expect AWP training to be highly effective, we do not present the method here due to the complexity of finding hyperparameters and the immense computational cost incurred for each training run, particularly in the case of ImageNet-100. We observe that the proposed approach achieves significantly improved robust accuracy against GAMA-PGD attack and AutoAttack [8] on the MNIST dataset, compared to even multi-step defenses such as TRADES and PGD-AT. We further remark that NuAT achieves this without a drop in standard accuracy on clean samples. We also observe again that NuAT achieves enhanced GAMA-PGD and AutoAttack accuracy as compared to all other defense methods on the highly challenging ImageNet-100 dataset, while requiring significantly less training time, demonstrating the scalability of the proposed NuAT defense. Lastly, we remark that exponential weight-averaging provides an additional boost in robust performance on ImageNet-100, on all three white-box attacks.

6.2 Computational Complexity

We compare the computational complexity of the proposed defense with various other efficient and effective defenses on the WideResNet-34-10 architecture for CIFAR-10 dataset in Fig.3(b). We report accuracy against the latest version (v2) of AutoAttack, which constitutes one of the strongest known ensemble of attacks till date. While the fastest defense amongst all is FBF [29], it achieves a significantly lower robust accuracy when compared to the other defenses. On the other extreme is AWP [30], which has the highest robust accuracy, but comes with a very high computational cost ($8\times$ that of NuAT2-WA) . The proposed defense NuAT2-WA achieves a robust accuracy of 54.76%, outperforming all other defenses (even TRADES, PGD) at a significantly lower computational cost.

6.3 Adaptive switching between single-step and two-step NuAT

In this work we propose two main variants of the NuAT defense, single-step and two-step. While single-step training results in the most effective defenses at low compute, two-step training is more stable, and yields results which are close to even the state-of-the-art multi-step (10-11 step) training methods. As observed in prior works [29, 18], single-step defenses cannot be used for long training regimes due to issues relating to stability of training. While the proposed NuAT single-step defense is significantly more stable compared to such works [29], it is not as stable as the two-step defense for large model capacities and longer training schedules. This limitation in the number of training epochs curtails the optimal performance that can be achieved by the base algorithm. Li et al. [18] use a validation set to detect the onset of instability in training and switch to a 10-step defense for a few iterations. Inspired by this, we propose to adaptively switch between the proposed single-step and two-step NuAT defenses by tracking a sudden drop in the training loss on clean samples when compared to the average loss of the previous epoch. While Li et al. [18] required a separate validation set and 10-step training, we seamlessly switch between our single-step and two-step defenses without any additional overheads. By varying the threshold at which we switch from single-step to two-step attacks, this method can either be used for improving the stability of the single-step run, or it can be used to boost the performance by switching to stronger two-step attacks when the model starts seeing early signs of gradient masking. For the hybrid NuAT-H and NuAT-WA-H models so trained, the additional epochs incurred are presented as a subscript in Table-1b. We thereby obtain significantly more robust models, while requiring an additional overhead of only upto 5 training epochs. We further expound upon adaptive switching integrated with NuAT in the Supplementary Section.

6.4 Sanity Checks to Verify the Absence of Gradient Masking

We further present detailed evaluations on several other targeted and untargeted white-box attacks, black-box attacks and gradient-free attacks in the Supplementary Section to verify the robustness of the single-step NuAT defense over all three datasets. Furthermore, we present also adaptive attacks that are cognizant of the proposed defense and other sanity checks in the Supplementary Material to verify the absence of gradient masking in the proposed defense as laid out by Athalye et al. [2].

7 Conclusions

In this work, we propose a novel Nuclear Norm regularizer to improve the adversarial robustness of Deep Networks through the use of single-step adversarial training. Training with the proposed Nuclear Norm regulariser enforces function smoothing in the vicinity of clean samples by incorporating joint batch-statistics of adversarial samples, thereby resulting in enhanced robustness. We demonstrate that the proposed NuAT defense out-performs existing single-step adversarial training methods, and scales well to large model capacities and large-scale datasets such as ImageNet-100. Additionally, we integrate exponential weight-averaging to achieve significantly improved robust performance. We further introduce a 2-step variant of the proposed defense that helps incorporate supervision from the weight-averaged model during training, resulting in models that achieve near state-of-the-art adversarial robustness at a significantly lower computational cost. The proposed defense method facilitates the effective training of robust networks through efficient means, and thus has immense potential to have a positive impact on society. We are also hopeful that future work could extend the proposed approach to other adversarial threat models of interest, improve upon the robustness-accuracy tradeoff incurred, and present provable robustness certificates to further justify worst-case performance.

References

[1] M. Andriushchenko, F. Croce, N. Flammarion, and M. Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *The European Conference on Computer Vision (ECCV)*, 2020. 7

[2] A. Athalye, N. Carlini, and D. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning (ICML)*, 2018. 1, 9

[3] J. Buckman, A. Roy, C. Raffel, and I. Goodfellow. Thermometer encoding: One hot way to resist adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2018. 1

[4] E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010. 3

[5] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017. 1

[6] T. Chen, Z. Zhang, S. Liu, S. Chang, and Z. Wang. Robust overfitting may be mitigated by properly learned smoothening. 7

[7] F. Croce and M. Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. In *International Conference on Machine Learning (ICML)*, 2020. 7

[8] F. Croce and M. Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning (ICML)*, 2020. 7, 8

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 7

[10] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015. 1, 3, 7

[11] S. Gowal, C. Qin, J. Uesato, T. Mann, and P. Kohli. Uncovering the limits of adversarial training against norm-bounded adversarial examples. *arXiv preprint arXiv:2010.03593*, 2020. 3

[12] C. Guo, M. Rana, M. Cisse, and L. van der Maaten. Countering adversarial images using input transformations. In *International Conference on Learning Representations (ICLR)*, 2018. 1

[13] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. G. Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018. 3, 7

[14] H. Ji, C. Liu, Z. Shen, and Y. Xu. Robust video denoising using low rank matrix completion. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1791–1798. IEEE, 2010. 3

[15] A. Krizhevsky et al. Learning multiple layers of features from tiny images. 2009. 7

[16] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial machine learning at scale. In *International Conference on Learning Representations (ICLR)*, 2017. 3

[17] Y. LeCun. The mnist database of handwritten digits. *Technical report*, 1998. 7

[18] B. Li, S. Wang, S. Jana, and L. Carin. Towards understanding fast adversarial training. *arXiv preprint arXiv:2006.03089*, 2020. 9

[19] A. Madry, A. Makelov, L. Schmidt, T. Dimitris, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018. 1, 2, 6, 7, 8

[20] T. Pang, X. Yang, Y. Dong, H. Su, and J. Zhu. Bag of tricks for adversarial training. *arXiv preprint arXiv:2010.00467*, 2020. 3, 7, 8

[21] T. K. Pong, P. Tseng, S. Ji, and J. Ye. Trace norm regularization: Reformulations, algorithms, and multi-task learning. *SIAM Journal on Optimization*, 20(6):3465–3489, 2010. 3

[22] L. Rice, E. Wong, and J. Z. Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning (ICML)*, 2020. 3, 7, 8

[23] A. Shafahi, M. Najibi, M. A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein. Adversarial training for free! In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 3, 7

[24] Y. Song, T. Kim, S. Nowozin, S. Ermon, and N. Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2018. 1

[25] G. Sriramanan, S. Addepalli, A. Baburaj, and R. Venkatesh Babu. Guided Adversarial Attack for Evaluating and Enhancing Adversarial Defenses. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2, 4, 5, 7, 8

[26] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2013. 1

[27] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations (ICLR)*, 2018. 4, 7, 8

[28] B. Vivek, A. Baburaj, and R. Venkatesh Babu. Regularizer to mitigate gradient masking effect during single-step adversarial training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019. 2, 4, 7, 8

[29] E. Wong, L. Rice, and J. Z. Kolter. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations (ICLR)*, 2020. 2, 4, 7, 8, 9

[30] D. Wu, S.-T. Xia, and Y. Wang. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems*, 33, 2020. 3, 7, 8, 9

[31] C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. Yuille. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations (ICLR)*, 2018. 1

[32] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning (ICML)*, 2019. 2, 7, 8

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Sections 5, 6
   (b) Did you describe the limitations of your work? [Yes] See Sections 5, 6, 7
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 7
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Will be included in the Supplementary Material
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 5, further details will be included in the Supplementary Material
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Section 6
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Will be included in the Supplementary Material
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 6, further details will be included in the Supplementary Material
   (b) Did you mention the license of the assets? [Yes] Details will be included in the Supplementary Material
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Will be included in the Supplementary Material
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]