Toward Efficient Robust Training against Union of ℓ_p Threat Models

Gaurang Sriramanan [1] **Maharshi Gor** [1] **Soheil Feizi** [1]

Abstract

The overwhelming vulnerability of deep neural networks to carefully crafted perturbations known as adversarial attacks has led to the development of various training techniques to produce robust models. While the primary focus of existing approaches has been directed toward addressing the worst-case performance achieved under a single-threat model, it is imperative that safety-critical systems are robust with respect to multiple threat models simultaneously. Existing approaches that address worst-case performance under the union of such threat models (e.g., $\ell_\infty, \ell_2, \ell_1$) either utilize adversarial training methods that require multi-step attacks which are computationally expensive in practice, or rely upon fine-tuning of pre-trained models that are robust with respect to a single-threat model. In this work, we show that by carefully choosing the objective function used for robust training, it is possible to achieve similar, or even improved worst-case performance over a union of threat models while utilizing only single-step attacks during the training, thereby achieving a significant reduction in computational resources necessary for training. Furthermore, prior work showed that adversarial training against the ℓ_1 threat model is relatively difficult, to the extent that even multi-step adversarially trained models were shown to be prone to gradient-masking and catastrophic overfitting. However, our proposed method—when applied on the ℓ_1 threat model specifically—enables us to obtain the first ℓ_1 robust model trained solely with single-step adversarial attacks.

1. Introduction

Recent years have demonstrated the success of deep learning in solving machine learning tasks spanning across various domains—computer vision, natural language texts, speech, etc. In addition, it has even exceeded the human level performance for certain tasks (He et al., 2016; 2015). However, despite their successes, these systems exhibit severe vulnerabilities: Deep learning models are very susceptible to imperceptible perturbations in the input at test time (Szegedy et al., 2013). Such human-imperceptible noise, known as adversarial attacks, can be used to induce networks to confidently predict incorrect labels, and can thus have disastrous implication in safety critical applications such as autonomous navigation and identity verification. To make models robust against such vulnerabilities at test time, a paradigm of *adversarial robust training* of machine learning models has been developed in recent years (Goodfellow et al., 2015; Madry et al., 2018; Zhang et al., 2019). These adversarial training procedures have primarily been used to train models robust to a single threat model—perturbations constrained within an ℓ_p-ball of ε_p radius for some p. For instance, the predominant threat model of interest that has been extensively studied in existing literature corresponds to the ℓ_∞threat model (mostly $\varepsilon_\infty = 8/255$). However, human-imperceptible adversarial perturbations can be sourced from multiple threat-models; hence in practice, it is pertinent to ensure that networks are robust against perturbations from a union of threat models simultaneously. More so, it has been observed that robust training procedures for a chosen threat model are not effective against attacks from other threat models (Tramer & Boneh, 2019; Maini et al., 2020), thus necessitating the development of adversarial defenses against multiple perturbation models simultaneously.

Over recent years, training procedures have been proposed to make systems simultaneously robust against perturbations constrained within a union of ℓ_∞, ℓ_1 and ℓ_2 balls. Systems trained in such manner are then evaluated over the worst-case performance across perturbations from all the threat-models. Tramer & Boneh (2019) proposed simple aggregations of different adversaries for adversarial training against multiple perturbation models utilizing multi-step adversarial attacks for robust training. Maini et al. (2020) further established SOTA for adversarial accuracy against union of (ℓ_∞, ℓ_1, ℓ_2) perturbations through the adversarial training procedure Multi Steepest Descent (or, MSD) that also uses multi-step ($k = 50$) adversarial attacks to generate adversaries for training. However, these methods, owing to their requirement of great number of adversarial training steps as compared to a regular setting for multi-step adversarial training procedure (10 steps), are computationally inefficient. This leads to our research question: Is it possible to achieve worst-case performance over a union of threat models that is similar to that of the SOTA methods, while

[1]Department of Computer Science, University of Maryland, College Park, USA. Correspondence to: Gaurang Sriramanan <gaurangs@cs.umd.edu>, Maharshi Gor <mgor@cs.umd.edu>, Soheil Feizi <sfeizi@cs.umd.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

utilizing training procedures that requires only single-step attacks to generate adversaries? We answer the same in affirmation: we first analyse failure modes of existing approaches during ℓ_1 based adversarial training, and thereby propose to use a dynamic curriculum schedule to effectively mitigate robust overfitting. Furthermore, we extend this approach to develop a training routine that utilizes a single-step adversarial training across a union of threat models to be robust against them simultaneously. In summary, we make the following contributions[1] in this work:

- We demonstrate the first successful single-step robust training procedure, NCAT-ℓ_1, to achieve ℓ_1 robustness by using a curriculum schedule with Nuclear Norm based training.
- We extend this approach to propose a training procedure NCAT, that yields SOTA-like robust accuracy under the union of multiple ℓ_p threat models, while requiring only a single-step attack budget per minibatch.
- We further demonstrate that the proposed defense can scale-up to high-capacity networks and large-scale datasets such as ImageNet-100. Additionally, NCAT trained models generalize to unseen threat models, achieving near-SOTA robustness even on Perceptual Projected Gradient Descent (PPGD), which comprises one of the strongest attacks known to date.

2. Preliminaries

Here, we lay down the notations and conventions used in this work. We denote x to be a d-dimensional image from an N-class dataset $\mathcal{D}$, while its corresponding ground-truth label as a one-hot vector y. f_θ represents 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 flattened pixel intensities of the image x_i, and Y as the corresponding ground-truth array. Thus, X is a matrix of size $(M \times d)$, and Y is a matrix of size $(M \times N)$. Also, $\ell_{CE}(f_\theta(X), Y)$ now denotes 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, the sum of the singular values, of A.

Adversarial Threat Model: In this work, we primarily consider the robustness of Deep Networks against the union of ℓ_∞, ℓ_1, and ℓ_2 constrained adversaries. Thus, a network f_θ is said to be ε_p-robust under a threat model ℓ_p on a clean sample x with label y, if $f_\theta(\tilde{x}) = y$, for all perturbations $\tilde{x}$ such that $\|x - \tilde{x}\|_p \leq \varepsilon_p$.

[1]Our code and pre-trained models are available here: https://github.com/GaurangSriramanan/NCAT.

3. Related Works

In this section we briefly discuss the adversarial attacks and defences that builds up to efficient multi-step adversarial training procedures, work that introduces adversarial training against the union of multiple threat models, and their limitations that we propose to alleviate.

While adversarial training methods have been observed to be the most effective defenses in recent times, early attempts of improving robustness to adversarial attacks included input pre-processing based defenses that were computationally cheap. However, such methods primarily relied upon masking of input gradients in order to counter white-box attacks. Several such defenses of this category were circumvented using smooth approximations of the non-differentiable components, or by utilizing expectation over randomized components (Athalye et al., 2018; Carlini et al., 2019).

3.1. Effectiveness of FGSM and its limitations

Perhaps the most successful defense which has stood the test of time is Projected Gradient Descent or PGD adversarial training (Madry et al., 2018). 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 standard training. Multi-step defenses achieve the state-of-the-art robustness today and typically use 10-steps of optimization for attack generation, leading to 11 times higher forward and backward propagations. FGSM or the Fast Gradient Sign Method (Goodfellow et al., 2015) based adversarial training alleviates the computational cost by utilizing single-step adversarial samples for training. However, in practice it is observed that during the course of FGSM training, degenerate solutions are frequently observed, wherein the local linearity assumption of the loss surface is violated. Indeed, Kurakin et al. (2017) showed that such models exhibited the phenomenon of gradient masking, wherein stronger multi-step attacks were seen to reduce the robust accuracy drastically. Wong et al. (2020) proposed to incorporate early-stopping using R-FGSM based adversarial training (Tramèr et al., 2018), in order to identify the failure-point during robust training with single-step adversaries. However, the method was later shown to not be effective on large capacity networks such as the WideResNet (Zagoruyko & Komodakis, 2016) architecture in subsequent work (Sriramanan et al., 2020).

Nuclear Norm Adversarial Training (NuAT): Sriramanan et al. (2021) proposes a Nuclear Norm regularizer to improve the adversarial robustness of Deep Networks through the use of single-step adversarial training under ℓ_∞constraints. This Nuclear Norm Adversarial Training (NuAT) enforces function smoothing in the vicinity of clean samples by incorporating joint batch-statistics of adversarial

samples, which results in enhanced robustness. Further, this limits the oscillation of function values and prevents the over-smoothing of loss surface uniformly in all directions, leading to a better robustness-accuracy trade-off.

3.2. Union of Threat Models

While above described works trains a target to be robust against a single threat model of ℓ_p-ball, there has been recent effort in the direction of making models robust against multiple threat models simultaneously. Tramer & Boneh (2019) study the theoretical and empirical trade-offs of adversarial robustness in various settings when defending against aggregations of multiple adversaries, proposing to train on the average (AVG) or maximizers of loss (MAX) amongst the different threat models considered for each minibatch of samples. Madaan et al. (2020) train using perturbations generated using a Meta-Noise Generator, and also propose a variant, Stochastic Adversarial Training wherein they utilize multi-step adversaries (10 steps for $\ell_\infty$and ℓ_2, 20 steps for ℓ_1), though the authors note sub-optimal performance from the same. Croce & Hein (2019) propose a provable adversarial defense against all ℓ_p norms for $p \geq 1$ using a regularization term for ReLU networks, by enforcing robustness against $\ell_\infty$and ℓ_1 adversaries in particular.

Multi Steepest Descent (MSD): The core idea that MSD (Maini et al., 2020) adopts, which helps establish better worst-case accuracies against the union of adversaries, is to create a single adversarial perturbation by simultaneously maximizing the worst-case loss over all perturbation models at each projected steepest descent step. Unlike previous approaches Tramer & Boneh (2019) that generate worst-case adversaries for each threat model, or augment adversaries from multiple threat models, MSD chooses a projected steepest descent direction in each iteration that maximizes the loss over all threat models. This has been established to be superior to the standard adversarial training and the simpler approaches that use comparatively myopic PGD subroutines that only use one perturbation model at a time. However, MSD requires 50 adversarial attack steps for each training iteration. Additionally, for each training step, it performs three forward passes (one for each threat model) and a backward pass.

Extreme Norms Adversarial Training (EAT): In order to achieve robustness against a union of ℓ_p threat models, Croce & Hein (2021b) propose to fine-tune models that were originally trained to be robust against a single ℓ_p norm threat model. The authors demonstrate that fine-tuning of robust models to previously unseen ℓ_p threat models is effective, in contrast to adversarial fine-tuning of normally trained networks which yields non-robust models. Furthermore, the authors propose to train solely on ℓ_1 and ℓ_∞adversaries, such that other ℓ_p balls on interest are contained within the union of these two threat models (Extreme Norms Adversarial Training or EAT). However, this can place excessive restrictions during robust training if the perturbation budget of intermediate ℓ_p adversaries is large. As with MSD (Maini et al., 2020), EAT is computationally expensive in practice, since it relies upon multi-step adversarially pre-trained models, and further performs robust fine-tuning of such models using 10-step adversaries in the second phase.

4. Proposed Method

As noted by prior works (Madry et al., 2018; Croce & Hein, 2021a; Maini et al., 2020), robust training against the ℓ_1 threat model is significantly more complicated when compared to standard adversarial training techniques for ℓ_∞or ℓ_2 threat models. Croce & Hein (2021a) note that even adversarial training using expensive 10-step adversaries generated from SLIDE (Tramer & Boneh, 2019) is prone to catastrophic overfitting (Wang et al., 2020): Over the course of training, models overfit to the adversaries generated, leading to a false notion of being robust, while achieving close to 0% accuracy against stronger attacks during test evaluation. While such phenomena are frequently seen in single-step training Goodfellow et al. (2015), the occurrence of such failure modes even with 10-step adversaries exhibits the difficulty in training ℓ_1 robust networks. Croce & Hein (2021a) demonstrate that using the 10-step APGD ℓ_1 attack, robust models can be trained by automatically tuning the sparsity level induced in the ℓ_1 perturbations seen during training. Building upon these in this work, we demonstrate the first successful instance of achieving ℓ_1 robustness using single-step adversaries during training. Further, we extend the technique to achieve simultaneous robustness against the union of ℓ_p threat models using single-step training.

4.1. Nuclear Norm Attack and Curriculum Schedule

We first focus on understanding the phenomenon of catastrophic overfitting under the ℓ_1 adversarial training and analyze what methods can help alleviate it in the single-step setting. We begin by plotting the prediction accuracy and cross-entropy loss of different models over training and validation (Figure 1). We find that when trained with R-FGSM based adversaries, models suffer from catastrophic overfitting early on during the training. However, we make a crucial observation that dynamically varying the perturbation budget during training, effectively setting up a *curriculum*, helps immensely in improving overall stability of training. For instance, with the final ℓ_1 threat model of interest given by the ball of radius 12, we propose to linearly increase this parameter from 0 to 12 to prevent catastrophic overfitting. However, applying this curriculum to RFGSM-AT only leads to a delay in catastrophic failure, indicating the unsuitability of using R-FGSM adversaries for robust training.

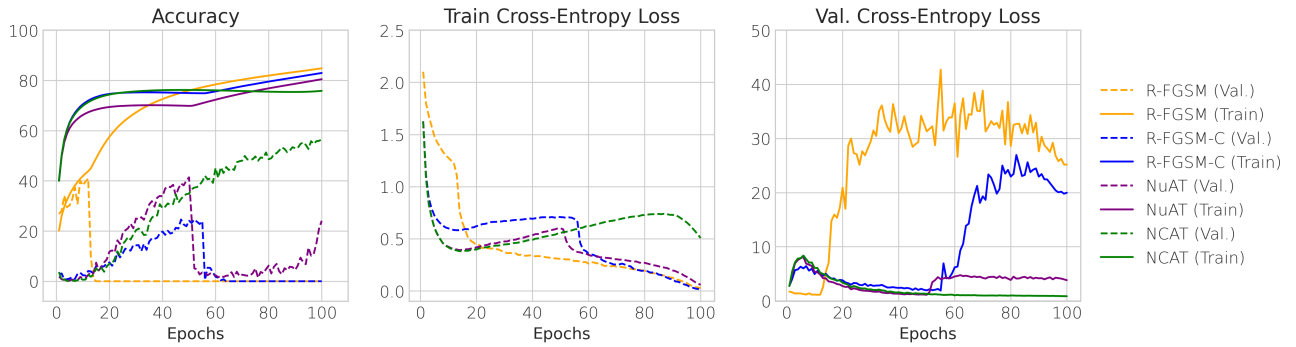

Figure 1: **Catastrophic Overfitting in ℓ_1 Adversarial Training:** To analyze the stability of single-step training, we plot accuracy (left) and cross-entropy losses (Right) over epochs of different single-step adversarially trained models. With R-FGSM based adversarial training (Wong et al., 2020), catastrophic overfitting occurs with extreme gradient masking (orange); adversarial accuracy (loss) is high (low) on the train set, while being close to zero (high) for validation images. More so, even using a curriculum schedule used for ℓ_1 adversaries during training only delays the overfitting (blue). In contrast, the proposed training approach NCAT (green) does not display catastrophic overfitting due to gradient masking.

Since the goal here is to utilize only single adversarial training step, it becomes imperative that the loss that we optimize over to generate the adversaries is smooth and does not showcase gradient masking. Hence, we build upon Nuclear-Norm Adversarial Training (NuAT) (Sriramanan et al., 2021) to generate single-step adversaries. Formally, in a given minibatch B, if X is the matrix composed of row-wise vectorized pixel values of each image, Δ is a matrix of independently sampled Bernoulli noise, and Y is the matrix containing the corresponding ground truth one-hot vectors, maximization of the following loss function that utilizes the pre-softmax values $f_\theta(\cdot)$ generates single-step adversaries:

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

Sriramanan et al. (2021) note that since the Nuclear norm forms a tight convex relaxation for the rank of the predicted matrix of logit values, the corresponding attack generates diverse adversaries in a given minibatch, which then helps mitigate robust overfitting. Crucially, we observe however that this supplemental attack diversity is not sufficient for single-step training on the ℓ_1 threat model, as even NuAT is observed to be susceptible to catastrophic failure in Fig-1. However, by utilizing the dynamic curriculum schedule, this phenomenon is successfully remedied in the proposed method, NCAT. Indeed, using this routine, we demonstrate for the first time that single-step training can be effectively used to produce ℓ_1 robust models.

4.2. Single-Step Training for ℓ_∞ and ℓ_2 Robustness

Since the ℓ_∞ threat model is the most well-studied setting in existing literature, we rely upon prior works to obtain excellent baselines. To achieve robustness against ℓ_∞ constrained adversaries using single-step training, we utilize the current state-of-the-art method, Nuclear Norm Adversarial Training (NuAT) (Sriramanan et al., 2021). We further seek to incorporate other threat models during training, in order to obtain models with non-trivial robustness against the union of the ℓ_1, ℓ_2 and ℓ_∞ threat models simultaneously. In order to efficiently train against ℓ_2 adversaries, we first propose to modify the NuAT training algorithm to utilize this constraint set, using ℓ_2 norm based projections. However, similar to Croce & Hein (2021b), we make the remarkable observation that models that are trained solely on ℓ_∞ adversaries achieve a great degree of robustness versus ℓ_2 adversaries on the test set. We observe similar transfer of robustness from ℓ_1 trained models toward the ℓ_2 threat model as well. Thus, the primary difficulty in achieving robustness to the union of threat models appears to be that of training networks robust to the ℓ_1 and ℓ_∞ threat models in particular.

4.3. Sampling Procedures to Improve Efficiency

To achieve robustness against the union of the three ℓ_p threat models considered, it is plausible that training with three distinct single-step attacks (constrained to ℓ_1, ℓ_2 and ℓ_∞) using the proposed approach in each minibatch will be effective. However, in this work, we primarily focus on reducing the training complexity further, in order to effectively utilise only a single-step attack for each minibatch. Awasthi et al. (2021) proposed to utilize the multiplicative weights algorithm, wherein the loss under different adversaries on a hold-out validation set guides the sampling procedure, using a set of exponential running weights w_i for each threat model. However, we find that this is contingent on the efficacy of adversaries utilized on the validation set, which can be restrictive in practice due to computational constraints. Building upon this, we find in practice that alternating be-

Algorithm 1 Nuclear Curriculum Adversarial Training for ℓ_p Norm Robustness

Input: Network f_θ with parameters θ, Weight Averaged Network f_ω with parameters ω, Training Data $\mathcal{D}$ with input images of dimension d, Minibatch Size M, Attack Size ε_p for each ℓ_p threat model, Epochs E, Learning Rate η, Decision Function D, Curriculum Schedule $\mathcal{C}$
for $epoch = 1$ **to** E **do**
  $\varepsilon_p = \mathcal{C}(p)$
  for minibatch $\{(x_i, y_i)\}_{i=1}^{M} \subset \mathcal{D}$ **do**
    $X = \begin{bmatrix} \dots & x_1 & \dots \\ \dots & \vdots & \dots \\ \dots & x_M & \dots \end{bmatrix}$, $\Delta = \begin{bmatrix} \dots & \delta_1 & \dots \\ \dots & \vdots & \dots \\ \dots & \delta_M & \dots \end{bmatrix}$
    $\delta_i \sim Bern^d(-\alpha, \alpha)$, $\widetilde{X} = X + \Delta$, $Y = \begin{bmatrix} y_1 \\ \vdots \\ y_M \end{bmatrix}$
    $\widetilde{L} = \ell_{CE}\left(f_\theta(\widetilde{X}), Y\right) + \lambda \cdot ||f_\theta(\widetilde{X}) - f_\theta(X)||_*$
    for p in $\mathrm{D}(\theta)$ **do**
      $\Delta = \Delta + \varepsilon_p \cdot \mathrm{Proj}\left(\nabla_\Delta \widetilde{L}, B_p(\varepsilon_p)\right)$
      $\widetilde{X} = Clamp\ (X + \Delta, 0, 1)$
    end for
    $L = \ell_{CE}(f_\theta(X), Y) + \lambda \cdot ||f_\theta(\widetilde{X}) - f_\theta(X)||_*$
    $\theta = \theta - \frac{1}{M} \cdot \eta \cdot \nabla_\theta L$
    $\omega = \tau \cdot \omega + (1 - \tau) \cdot \theta$
  end for
end for

tween ℓ_∞ and ℓ_1 attacks across different minibatches with a fixed frequency is remarkably effective. Thus, the proposed defense, NCAT, uses nuclear norm based single-step training following a curriculum schedule, such that different threat models are selected for attack generation in different minibatches based on a pre-fixed frequency. We present a concise, summarised overview of the proposed training approaches in Algorithm-1. Here, the Decision Function D (L8, Alg-1) alternately outputs $p = 1$ or $p = \infty$ based on a predetermined frequency, since such models are observed to simultaneously achieve ℓ_2 robustness without explicit training. As observed in prior works (Chen et al.; Sriramanan et al., 2021), maintaining a exponential running average of network weights (SWA (Izmailov et al., 2018)) helps improve robust performance overall as well, particularly so in this setup since different (random) minibatches are trained with adversarial perturbations arising from different threat models. Furthermore, this effectively reducing undesired bias to a particular threat model due to auto-correlations that arise in training.

5. Experiments and Analysis

In this work, we primarily consider the CIFAR-10 (Krizhevsky et al., 2009) and ImageNet-100 (Russakovsky et al., 2014) datasets, since they have come to form the benchmark for comparative analysis of adversarially robust models. Following prior works (Maini et al., 2020), we consider constraint sets given by the ℓ_∞ball of radius $8/255$, ℓ_2 ball of radius 0.5 and ℓ_1 ball of radius 12 as the threat models of interest, and as explained previously, we attempt to train models that achieve non-trivial worst-case accuracy against the union of such ℓ_p threat models. For the ImageNet-100 dataset, the corresponding radii for ℓ_1, ℓ_2 and ℓ_∞threat models are 255, 1200/255 and 4/255 respectively, following the constraints considered by Laidlaw et al. (2021). We present results in the white-box setting, wherein the adversary is cognizant of the model weights, architecture and training scheme employed. To accurately estimate worst-case performance, we focus our evaluation pipeline to incorporate state-of-the-art attacks such as AutoAttack (Croce & Hein, 2020) for each ℓ_p threat model. Furthermore, AutoAttack includes strong ℓ_1 attack evaluation baselines using techniques proposed by Croce & Hein (2021a), wherein the authors note that significant improvement in attack efficacy as compared to prior works. We further include black-box evaluations, generalization to unseen domains, gradient masking checks and adaptive attacks in Sections-A1, A2, A3 of the Appendix.

We first present results obtained using the ResNet-18 (He et al., 2016) architecture on CIFAR-10 in Table-1. In the first partition of the table, we present models trained solely on the ℓ_1 threat model. The current state-of-the-art is achieved by APGD-ℓ_1 (Croce & Hein, 2021a), which performs a 10-step APGD attack during training in order to mitigate gradient masking and catastrophic overfitting for ℓ_1 constrained adversaries. On the other hand, our method, NCAT-ℓ_1 which uses just a single-step attack for adversarial training achieves ℓ_1 robustness much more efficiently. We note that while the multi-step approach has higher ℓ_1 robustness (+3.6%), the single-step NCAT trained model has significantly better worst-case accuracy over all three threat models (+15.8). Indeed, we note once again that NCAT represents the first-ever successful single-step adversarial training on the ℓ_1 threat model, which also generalizes well to the unseen $\ell_\infty$and ℓ_2 threat models simultaneously.

In the second partition of Table-1, we present models that are explicitly trained to be robust under the union of the ℓ_∞, ℓ_2 and ℓ_1 threat models. Namely, we present comparative analysis with respect to existing multi-step adversarially trained defenses such as AVG and MAX (Tramer & Boneh, 2019), MSD (Maini et al., 2020), EAT (Croce & Hein, 2021b) and SAT (Madaan et al., 2020). For these methods, we primarily utilize robust evaluations as presented by Croce &

Table 1: **CIFAR-10:** Prediction accuracy (%) of ResNet-18 models trained using different methods under various threat models. Robust evaluations are presented under the constraint sets given by $\varepsilon_1 = 12$, $\varepsilon_2 = 0.5$ and $\varepsilon_\infty = 8/255$ for individual threat models, along with worst-case and average-case performance under their union.

Method	Number of AT Steps	Clean Acc	Worst-Case Acc	Average Acc	ℓ_1 Acc	ℓ_2 Acc	ℓ_∞ Acc
ℓ_1 Training Alone							
APGD-ℓ_1	10	85.9	22.1	48.8	59.5	64.9	22.1
NCAT-ℓ_1	1	80.6	36.8	53.2	55.5	67.4	36.8
Training under Union of Threat Models							
SAT	13.33†	83.9	40.4	54.2	54.0	68.0	40.7
AVG	30	84.6	40.1	53.8	52.1	68.4	40.8
MAX	30	80.4	44.0	53.4	48.6	66.0	45.7
MSD	50	81.1	43.9	53.4	49.5	65.9	44.9
EAT	10††	82.2	42.4	54.6	53.6	67.5	42.7
NCAT	1	80.3	42.6	53.3	46.9	67.0	46.0
NCAT$^+$	1	77.5	43.7	53.4	48.4	65.7	46.1

Hein (2021b) to enable fair comparisons, which comprise of re-implemented models that obtain higher accuracies as compared to values reported in the original papers. We first note that SAT† requires 13.33 adversarial attack steps during training, since it utilizes 10-step attacks for $\ell_\infty$and ℓ_2 adversaries, and 20 attack steps for ℓ_1 adversaries to mitigate gradient masking, indicating the considerable difficulty involved in achieving ℓ_1 robustness. In contrast, EAT†† relies upon 10-step fine-tuning of a network that is already robust against a single threat-model. The current state-of-the-art approaches comprise of MSD and MAX that achieve 44% worst-case accuracy, while utilizing a budget of 50 and 30 attack steps respectively during training. We observe that the proposed approach, NCAT achieves comparable worst-case and average-case performance over the threat models considered, while requiring a significantly smaller computational footprint during training. Furthermore, the proposed method facilitates the trade-off between clean accuracy and worst-case performance as indicated by NCAT$^+$ which achieves near-SOTA robust performance.

In Table-2, we present results on models trained with the WideResNet-28-10 (Zagoruyko & Komodakis, 2016) architecture to demonstrate the scalability of the proposed defense to high-capacity networks. We thus establish the efficacy of the curriculum schedule combined with nuclear norm based training in mitigating catastrophic overfitting, enabling efficient training of these large networks.

In Table-3, we present evaluations on the ImageNet-100 dataset, wherein we utilize ResNet-18 networks to reduce computational demands. We observe that NCAT-ℓ_1 attains remarkable robust accuracy on unseen $\ell_\infty$and ℓ_2 adversaries, even achieving 26.6% accuracy against the Perceptual Projected Gradient Descent (PPGD) attack (Laidlaw et al., 2021), which forms one of the strongest attacks known to date. Further, we observe that the NCAT trained achieves

Table 2: **CIFAR-10:** Prediction accuracy (%) of WideResNet-28-10 models trained using different methods under various threat models. Robust evaluations are presented under the constraint sets given by $\varepsilon_1 = 12$, $\varepsilon_2 = 0.5$ and $\varepsilon_\infty = 8/255$ for individual threat models, along with worst-case and average-case performance under their union.

Method	Number of AT Steps	Clean Acc	Worst-Case Acc	Average Acc	ℓ_1 Acc	ℓ_2 Acc	ℓ_∞ Acc
ℓ_1 Training Alone							
APGD-ℓ_1	10	83.7	30.7	52.5	61.6	65.1	30.7
NCAT-ℓ_1	1	80.7	39.2	54.6	56.1	68.6	39.3
Training under Union of Threat Models							
SAT	13.33†	80.5	45.7	56.2	55.9	66.7	45.9
AVG	30	82.5	45.1	56.1	55.0	68.0	45.4
MAX	30	79.9	47.4	54.6	50.2	65.3	48.4
MSD	50	80.6	46.9	55.1	51.7	65.6	48.0
EAT	10††	79.9	46.4	56.3	56.0	66.2	46.6
NCAT	1	81.5	44.6	54.8	49.9	68.3	46.3

Table 3: **ImageNet-100:** Prediction accuracy (%) of models trained using different methods under various threat models. Robust evaluations are presented under the constraint sets given by $\varepsilon_1 = 255$, $\varepsilon_2 = 1200/255$ and $\varepsilon_\infty = 4/255$ for individual threat models, along with worst-case and average-case performance under their union.

Method	Number of AT Steps	Arch	Clean Acc	Worst-Case Acc	Average Acc	ℓ_1 Acc	ℓ_2 Acc	ℓ_∞ Acc	PPGD Acc
ℓ_∞-AT	10	RN50	81.7	0.8	20.7	0.8	3.7	55.7	1.5
PAT	10	RN50	72.6	37.8	41.2	41.2	37.7	45.0	29.2
NCAT-ℓ_1	1	RN18	64.9	41.1	43.9	48.3	41.4	42.1	26.6
NCAT	1	RN18	63.9	44.6	43.9	46.8	41.9	45.7	29.1

state-of-the-art worst-case ℓ_p accuracy, while attaining robustness similiar to that of Perceptual Adversarial Training (Laidlaw et al., 2021) for the PPGD attack, with the latter being a model that was explicitly trained on such adversaries.

6. Conclusions

In this work, we develop an efficient adversarial training procedure, NCAT, to train networks that are robust against a union of ℓ_p threat models, namely ℓ_∞, ℓ_1 and ℓ_2. To do so, we first focus on developing an efficient, yet effective robust training procedure for the ℓ_1 threat model, by incorporating a curriculum schedule to mitigate catastrophic overfitting. Indeed, in this work we present the first ℓ_1 constrained robust model trained solely using single-step adversaries, achieving robustness similar to that of multi-step SOTA approaches. Furthermore, we extend the proposed method to achieve worst-case robustness under multiple ℓ_p norm constraints simultaneously. Compared to the current SOTA that uses 30 adversarial attack steps for its training procedure to achieve 44% robust accuracy on CIFAR-10, our method yields 43.7% robustness while solely utilizing single-step adversaries during the training routine. This thereby greatly reduces the computational requirements needed to achieve SOTA-equivalent robust performance.

References

Andriushchenko, M., Croce, F., Flammarion, N., and Hein, M. Square attack: a query-efficient black-box adversarial attack via random search. In *The European Conference on Computer Vision (ECCV)*, 2020. 9

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

Awasthi, P., Yu, G., Ferng, C.-S., Tomkins, A., and Juan, D.-C. Adversarial robustness across representation spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7608–7616, 2021. 4, 12

Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., Goodfellow, I., and Madry, A. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019. 2

Chen, T., Zhang, Z., Liu, S., Chang, S., and Wang, Z. Robust overfitting may be mitigated by properly learned smoothening. 5

Croce, F. and Hein, M. Provable robustness against all adversarial l_p-perturbations for $p \geq 1$. *arXiv preprint arXiv:1905.11213*, 2019. 3

Croce, F. and Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning (ICML)*, 2020. 5, 9, 11

Croce, F. and Hein, M. Mind the box: l_1-APGD for sparse adversarial attacks on image classifiers. 139:2201–2211, 2021a. 3, 5

Croce, F. and Hein, M. Adversarial robustness against multiple l_p-threat models at the price of one and how to quickly fine-tune robust models to another threat model. May 2021b. 3, 4, 5

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015. 1, 2, 3

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *arXiv preprint arXiv: Arxiv-1502.01852*, 2015. 1

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 5, 11

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm. 9

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., and Wilson, A. G. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018. 5, 11

Kang, D., Sun, Y., Hendrycks, D., Brown, T., and Steinhardt, J. Testing robustness against unforeseen adversaries. *arXiv preprint arXiv:1908.08016*, 2019. 9

Krizhevsky, A. et al. Learning multiple layers of features from tiny images. 2009. 5, 11

Kurakin, A., Goodfellow, I., and Bengio, S. Adversarial machine learning at scale. In *International Conference on Learning Representations (ICLR)*, 2017. 2

Laidlaw, C., Singla, S., and Feizi, S. Perceptual adversarial robustness: Defense against unseen threat models. In *ICLR*, 2021. 5, 6, 11

Madaan, D., Shin, J., and Hwang, S. J. Learning to generate noise for multi-attack robustness, 2020. URL https://arxiv.org/abs/2006.12135. 3, 5

Madry, A., Makelov, A., Schmidt, L., Dimitris, T., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018. 1, 2, 3, 13

Maini, P., Wong, E., and Kolter, J. Z. Adversarial robustness against the union of multiple perturbation models. In *International Conference on Machine Learning*, 2020. URL https://arxiv.org/abs/1909.04068. 1, 3, 5, 11

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019. 11

Rice, L., Wong, E., and Kolter, J. Z. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning (ICML)*, 2020. 13

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein,

M. S., Berg, A. C., and Fei-Fei, L. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014. URL http://arxiv.org/abs/1409.0575. 5, 11

Smith, L. N. No more pesky learning rate guessing games. *CoRR*, abs/1506.01186, 2015. URL http://arxiv.org/abs/1506.01186. 11

Sriramanan, G., Addepalli, S., Baburaj, A., and Venkatesh Babu, R. Guided Adversarial Attack for Evaluating and Enhancing Adversarial Defenses. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2, 10, 11

Sriramanan, G., Addepalli, S., Baburaj, A., and Radhakrishnan, V. B. Towards efficient and effective adversarial training. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=kuK2VARZGnI. 2, 4, 5, 11

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

Tramer, F. and Boneh, D. Adversarial training and robustness for multiple perturbations. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 3, 5

Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., and McDaniel, P. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations (ICLR)*, 2018. 2

Wang, Y., Zou, D., Yi, J., Bailey, J., Ma, X., and Gu, Q. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations (ICLR)*, 2020. 3

Wong, E., Rice, L., and Kolter, J. Z. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations (ICLR)*, 2020. 2, 4, 12

Zagoruyko, S. and Komodakis, N. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 2, 6, 11

Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., and Jordan, M. I. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning (ICML)*, 2019. 1, 13

Appendix

A1. Black-Box and Zeroth-Order Attacks

Table A1: **Black-Box and Unseen attacks on CIFAR-10:** Prediction accuracy (%) of ResNet-18 models trained using NCAT-ℓ_1 and NCAT.

Method	Number of AT Steps	Clean Acc	Square ℓ_1	Square ℓ_∞	Common Corr.	Elastic	Gabor
APGD-ℓ_1	10	87.1	71.8	40.8	72.0	48.7	12.4
NCAT-ℓ_1	1	81.7	65.2	48.5	67.0	54.1	12.9
NCAT	1	80.3	60.1	53.8	65.0	71.4	14.9

While white-box attacks that utilize first-order methods generally form the strongest suite of adversarial perturbations, it is plausible that models are not inherently robust, but rather rely upon obfuscated or shattered gradients (Athalye et al., 2018) to falsely display high robust accuracies against such attacks. In this section, we thus present robust evaluations using attack methods that do not rely upon gradient information to craft adversaries.

For Black-box evaluation, we primarily rely on the Square attack (Andriushchenko et al., 2020), since it has been shown to be the strongest gradient-free attack presently. As shown in Table-A1, the NCAT and NCAT-ℓ_1 models achieves significantly higher robust accuracy on the Square Attack as compared to the evaluation presented in Table-1 of the Main paper, indicating that zeroth-order adversaries are weaker than gradient-based attacks. As expected, we also note that the NCAT model trained explicitly on the union of threat models obtains higher Square ℓ_∞ accuracy as compared to the NCAT-ℓ_1 model. On the other hand, for Square ℓ_1 adversaries, the NCAT-ℓ_1 model outperforms NCAT by 5%, since training on specific adversaries on a narrow threat model is more efficacious against similar adversaries during test-time. Comparing with the APGD-ℓ_1 which takes 10 adversarial steps, our approach transfers significantly better over attacks from other threat models:NCAT-ℓ_1 performs roughy 8% better than APGD-ℓ_1 on Square Attack-ℓ_∞, though ℓ_1 specific robustness is lower as seen with the Square-ℓ_1 attack.

We further verify that such black-box adversaries are indeed weaker than the suite of white-box attacks presented in the main paper, thereby helping confirm the absence of obfuscated gradients in the proposed NCAT trained model.

A2. Generalization to Unseen Domains

In the right-hand partition of Table-A1, we present evaluations of the NCAT-ℓ_1 and NCAT trained models on domain shifts that are not seen during training. We observe that the single-step trained models generalize well to images with common corruptions, obtaining 67% and 65% on the CIFAR10-C dataset (Hendrycks & Dietterich, 2019) with the highest severity setting (5). The slight increase (0.2%) in the case of the NCAT-ℓ_1 is likely due to the base clean accuracy being higher as compared to the NCAT model. Similarly, the APGD-ℓ_1 trained model obtains higher accuracy on CIFAR10-C largely due to higher performance on clean samples. We also evaluate the model on Elastic and Gabor Transformations as introduced by Kang et al. (2019). For Elastic image distortions, the NCAT model performs significantly better (+17.4%) as compared to the NCAT-ℓ_1 which was trained solely against ℓ_1 adversaries. Further, we observe that single-step training with NCAT or NCAT-ℓ_1 achieves higher accuracy as compared to the APGD-ℓ_1 trained model, with even NCAT-ℓ_1 achieving an improvement of 5.4% on Elastic distortions over the latter, indicating the improved generalization seen with single-step training. However, for other distortions such as Gabor, prediction accuracy is significantly lower for all three models.

A3. Gradient Masking Checks and Adaptive Attacks

In order to verify that the white-box attacks utilized are indeed effective in identifying strong adversaries within the considered threat model of interest, we present more detailed robust evaluations (Athalye et al., 2018) for the proposed NCAT trained ResNet-18 model in Fig.-A1. Here, we present the accuracy versus epsilon plot, and cross-entropy loss versus epsilon plot for the NCAT-ℓ_1 model in the first column on ℓ_1 APGD-CE (Croce & Hein, 2020) adversaries. In the latter three columns, we present the same metrics on ℓ_1, ℓ_2 and ℓ_∞ APGD-CE attacks for various values of epsilon for the NCAT model trained to be robust against the union of such adversaries. In each case, we observe that the robust accuracy monotonically decreases to zero as the perturbation budget (ε) is increased. Further, the cross-entropy loss monotonically increases as the perturbation budget (ε) is increased. This shows that gradient-based white-box attacks are strong and effective, with a smooth local loss landscape, indicating the absence of gradient masking in the single-step defenses NCAT-ℓ_1 and NCAT.

Further, we evaluate the NCAT defense against adaptive adversaries that incorporate modified objectives to obtain stronger attacks, since we assume that adversaries are cognizant of the training methodology used. We thus maximise the Nuclear Norm objective, (Eq-1 of the Main paper) to

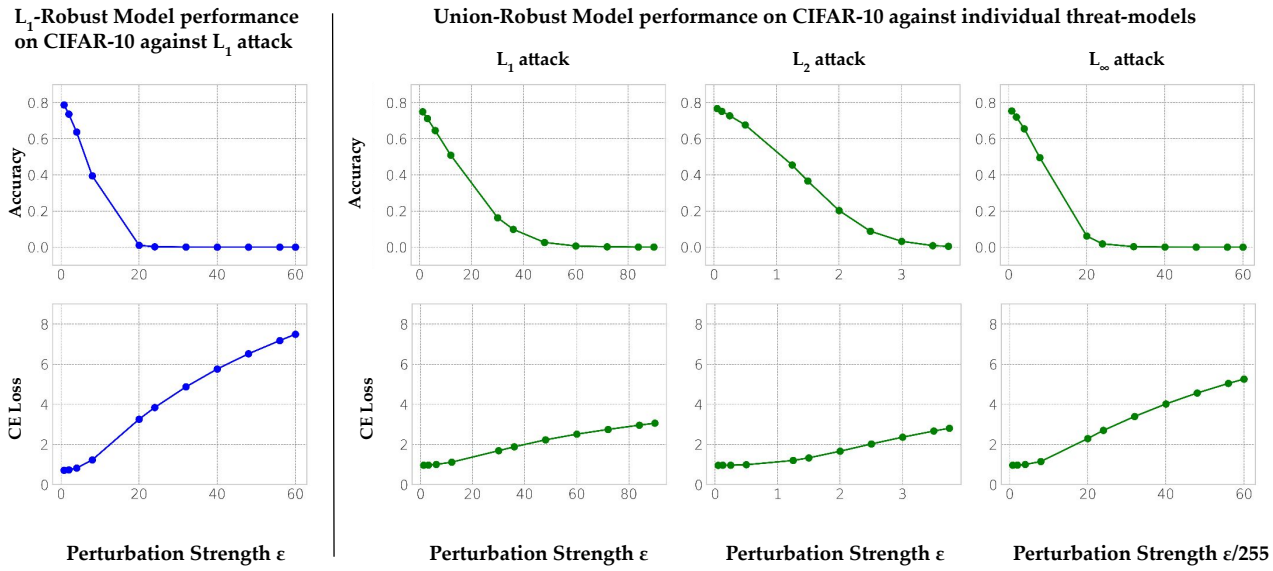

Figure A1: **Robustness across varying Perturbation Strengths** Row-1: Robust Accuracy is plotted for APGD-CE adversaries of different perturbation strengths for the NCAT-ℓ_1 model in the left partition, and for the NCAT model robust in the right partition. Row-2: Cross-Entropy Loss is plotted for APGD-CE adversaries of different perturbation strengths for the NCAT-ℓ_1 model in the left partition, and for the NCAT model robust in the right partition.

generate adaptive adversaries:

$$\widetilde{L} = \ell_{CE}\left(f_\theta(X+\Delta), Y\right) + \lambda \cdot ||f_\theta(X+\Delta) - f_\theta(X)||_* \quad \text{(A1)}$$

Since the AutoAttack framework utilizes automatic updates to the step-size with restarts at the iteration that maximizes the overall loss, the incorporation of the Nuclear norm regularizer is sub-optimal since batch-statistics across different images weaken the attack due to the reduced specificity in perturbations. We further implement an ℓ_1-version of GAMA-PGD (Sriramanan et al., 2020) to incorporate the Nuclear norm objective, with a decaying coefficient for the regularization term in order to mitigate this effect. However, we find that this adaptive adversary is weak once again, with NCAT achieving 75.3% accuracy. Thus, we find that the adaptive attacks are not stronger than the evaluations performed using AutoAttack as presented in the Main paper, and that the latter is sufficient to obtain a reliable estimate of the worst-case ℓ_1 accuracy obtained by the NCAT model.

A4. Steepest Ascent with Single-Step Optimization

As explained in the main paper, we generate the Nuclear Norm attack by identifying a perturbation that maximizes the loss $\widetilde{L}$ as in Eq-A1, such that it conforms to the ℓ_1 ball and [0,1] pixel-wise image constraints. Assuming a first-order Taylor series approximation for the loss incurred by the network f_θ, the steepest ascent direction to maximize loss would be parallel to the gradient direction in the unconstrained setting. Thus, if g represents the gradient of the loss for an image x, for steepest ascent of loss we have:

$$\max_{\delta} \left[\sum_{i=1}^{d} g_i \delta_i\right] \text{ such that}$$
$$\text{(a) } 0 \leq x_i + \delta_i \leq 1 \; \forall \; i, \text{ and (b) } ||\delta||_1 \leq \epsilon_1 \quad \text{(A2)}$$

In the absence of constraint (b), the optimal perturbation δ is given by $\delta = M$ where M_i denotes the deviation budget required at pixel i to saturate the same to the pixel constraint $[0, 1]$ parallel to the gradient g_i, and can defined formally as:

$$M_i = \begin{cases} 1 - x_i & \text{if } g_i \geq 0 \\ -x_i & \text{if } g_i < 0 \end{cases}$$

When the overall available budget is limited by ε_1 as in constraint (b), such that $||M||_1 < \varepsilon_1$, the constraint is inactive, and the solution is unaltered. On the other hand, if constraint (b) is active, the solution is necessarily different, wherein a reduction in the perturbation allocated for some pixel locations is made mandatory. Thus, inner-product in Eq-A2 is maximized by assigning the perturbations M_i in priority-order for different pixel locations, based on decreasing magnitude of the absolute gradient values. Let σ denote the sorted permutation of indices such that $a_{\sigma(1)} \geq a_{\sigma(1)} \geq \cdots \geq a_{\sigma(d)}$, where $a_i = |g_i|$ represents the gradient magnitudes at different pixel locations.

Further, let the cumulative budget utilized be defined as $S_i = \Sigma_{j=1}^{i} |M_{\sigma(j)}|$. Since each term $|M_{\sigma(j)}|$ in the summand is positive, S_i increases monotonically. Thus with $I_i = \max\{0, \varepsilon_1 - S_{i-1}\}$ denoting support variables for indices which receive a lower perturbation allocation due to constraint (b), the optimal single-step perturbation corresponding to image x is then defined as:

$$\delta^*_{\sigma(i)} = \begin{cases} M_{\sigma(i)} & \text{if } S_i \leq \varepsilon_1 \\ M_{\sigma(i)} \cdot I_i & \text{if } S_i > \varepsilon_1 \end{cases} \tag{A3}$$

Thus, in an ℓ_1 constrained attack the gradients have to be sorted by their magnitude, which requires $O(d \cdot \log d)$ complexity where d represents the image dimensionality. However, in practice this overhead is observed to be exceedingly minimal relative to data loading times etc.

A5. Implementation Details and Training Methodology

A5.1. Details on Datasets

In this work, we present our evaluations on the CIFAR-10 (Krizhevsky et al., 2009) and ImageNet-100 (Russakovsky et al., 2014) datasets, as they have come to form the benchmark datasets for robust evaluations.

CIFAR-10 (Krizhevsky et al., 2009) is a ten-class dataset, consisting of 32×32 sized RGB images arising from the following categories: "airplane", "automobile", "bird", "cat", "deer", "dog", "frog", "horse", "ship" and "truck". The test set of CIFAR-10 consists of 10,000 images, and the original training set consists of 50,000 images. The latter is split in practice, to form 49,000 training images and a hold-out validation set of 1000 images. On this dataset, we present robust evaluations against adversaries constrained under an ℓ_1 ball of radius 12, ℓ_2 ball of radius 0.5 and an ℓ_∞ ball of radius 8/255, similar to the setting considered in prior work (Maini et al., 2020; Croce & Hein, 2020).

ImageNet-100 is a hundred-class subset of the original ImageNet Large Scale Visual Recognition Challenge (Russakovsky et al., 2014), wherein every tenth class by WordNet ID order is retained, similar to the methodology followed by Laidlaw et al. (2021). This dataset consists of 224×224 sized RGB images arising from a diverse set of classes. Given the high-dimensional nature of the images, and the diversity of classes, it is quite challenging to train robust models effectively on this dataset. On this dataset, we present robust evaluations against adversaries constrained under an ℓ_1 ball of radius 255, ℓ_2 ball of radius 1200/255 and an ℓ_∞ ball of radius 4/255, similar to Laidlaw et al. (2021). Furthermore, images in this dataset are more realistic, with higher visual fidelity as compared to CIFAR-10. We thus present results on the unseen Neural Perceptual Threat Model (NPTM) (Laidlaw et al., 2021) on this dataset in Table-3 of the Main Paper using the Perceptual Projected Gradient Descent (PPGD) attack for the medium NPTM bound (0.5).

A5.2. Training and Hyperparameter Details

In this work, all training and experimental evaluations were performed using Pytorch (Paszke et al., 2019). We primarily utilize the ResNet-18 (He et al., 2016) architecture for both the CIFAR-10 and ImageNet-100 datasets. In addition, we present results on models trained on CIFAR-10 using the WideResNet-28-10 (Zagoruyko & Komodakis, 2016) architecture, that is, a WideResNet network with a depth of 28, and a width-factor of 10. We utilise a 100-epoch training schedule for the ResNet-18 models, and a 50-epoch regime for training WideResNet models. In all training runs, we use a cyclic schedule (Smith, 2015), with the maximum learning rate set to 0.1. We further utilize the Stochastic Gradient Descent (SGD) optimizer using a momentum parameter set to 0.9 and weight-decay of $5\text{e-}4$. Further, we utilize Random-Crop and Random-Horizontal-Flip as augmentations for training images. Similar to prior works (Izmailov et al., 2018; Sriramanan et al., 2021), we utilize Stochastic Weight Averaging with the exponential parameter τ being largely optimal, together with a setting of 0.9998 with a batch-size of 64. For NCAT-ℓ_1, we set the coefficient of the Nuclear Norm regularizer λ to 5, and for NCAT we use $\lambda = 3$ for ℓ_1 adversaries and $\lambda = 5$ for ℓ_∞ adversaries to achieve robustness against the union of threat models. The proposed approach NCAT requires the same computational complexity in training as Nuclear Norm Adversarial Training (NuAT), and thus achieve the same reduction in computational requirements over existing multi-step approaches as reported by Sriramanan et al. (2021). We use Nvidia RTX 2080 TI and Nvidia RTX A4000 GPU cards for training and experimental evaluations.

A5.3. Details on Curriculum Schedule

As explained in Section-4.1 of the Main Paper, we propose to utilize a curriculum schedule for training on adversarial perturbations of increasing difficulty over the training regime. To do so, we linearly increase the radius of the ℓ_p ball considered for generating adversaries, thereby significantly reducing the extent of overfitting and eliminating catastrophic failure entirely during training. Further, we linearly increase the coefficient of the Nuclear norm regularization term λ in-sync with the increase in ℓ_p radii. These techniques are particularly efficacious when we seek to achieve robustness against multiple threat models simultaneously, since different threat models can offer relatively different strengths of adversaries as the radii are increased during training. Similar to Sriramanan et al. (2020), we also set the value of λ used in the attack to zero in alternate mini-

batches, in order to further boost diversity of adversaries seen during the training regime. In practice, we require that the model is trained on the final union of threat models for a sufficiently short duration, comprising of a few epochs of training. Hence, we linearly ramp up the ℓ_p radii such that adversaries are generated from the final threat model of interest in the last 10 epochs of training, following which the radii are kept constant.

A6. Ablation Analysis

Table A2: **Ablations on CIFAR-10:** Prediction accuracy (%) of ResNet-18 models trained on the ℓ_1 threat model using NCAT-ℓ_1 (left), and on the union of ℓ_1, ℓ_2 and ℓ_∞ threat models using NCAT (right). Robust accuracy is computed using only ℓ_1 adversaries in the left partition, while worst-case accuracy over adversaries constrained under the union of ℓ_1, ℓ_2 and ℓ_∞ threat models is presented in the right partition.

Method	Clean Acc	ℓ_1 Robust Acc	Method	Clean Acc	Worst-Case Acc
A1: RFGSM-ℓ_1	89.9	0.0	A5: Exp. Wts. Samp	79.9	39.2
A2: RFGSM-ℓ_1 + Early-stop.	71.8	32.5	A6: NCAT-AVG	79.1	40.4
A3: NuAT-ℓ_1	92.8	1.2	A7: NCAT $p = 0.4$	80.9	42.4
A4: NuAT-ℓ_1 + Early-stop.	81.2	36.1	A8: NCAT $p = 0.6$	80.1	42.0
NCAT-ℓ_1	80.6	55.5	NCAT	80.5	42.5

In this section, we perform ablative experiments to study the significance of different components in the proposed defense. In the left partition of Table-A2, we present results for various ℓ_1 trained models, while the right partition corresponds to models that are trained to be robust against adversaries under the union of ℓ_1, ℓ_2 and ℓ_∞ threat models. In Ablations A1 and A2, we present results obtained using RFGSM training (Wong et al., 2020), wherein we note that catastrophic failure occurs early during the course of training. Even with early-stopping as suggested by Wong et al. (2020), the model obtains low clean accuracy (71.8%), and subpar robust accuracy due to the early collapse in training. We observe a similar phenomenon with Nuclear Norm adversarial training (A3,A4), wherein the model undergoes failure at a delayed phase as compared to RFGSM trained models. Thus, though NuAT obtains improved results, catastrophic failure during training results in the sub-par models with very low robust performance (36.1%). However, with the curriculum schedule as explained in Section-A5.3 and Section-4.1 of the Main Paper, the training dynamics in NCAT is highly stabilized, resulting in the first ℓ_1 robust model trained solely using single-step adversaries.

In the right partition, we first present ablation A5, wherein the frequency of sampling adversaries from different threat models is dynamically altered according to an exponential weights algorithm as proposed by Awasthi et al. (2021), based on metrics recorded on a hold-out validation set. In practice, these updates are seen to be excessively sensitive to the degree of convergence achieved by adversaries on the validation set resulting in lower robust accuracy on the union of adversaries (39.2%), and further requires additional hyperparameter tuning for the exponential weighting, along with an added computational budget for recording validation performance at each epoch. In Ablation A6, we present NCAT-AVG, which uses a Decision Function D that outputs the collection of $p = \{1, 2, \infty\}$ in Line-8 of Alg-1, and effectively uses a budget of three single-step attacks, one for each threat model. Further, we observe that the robust accuracy under the union of threat models is reduced despite the increase in training cost, and is accompanied with reduction in clean performance as well. Lastly, we present ablations A7 and A8 where the frequency of sampling ℓ_∞ based adversaries is changed to $p = 0.4, p = 0.6$ respectively. In practice, it is highly plausible that a subset of specified threat models is significantly simpler to achieve robustness as compared to other adversaries. This sampling mechanism helps incorporate the same in a simple manner, and subsumes NCAT which utilizes $p = 0.5$ for all experiments. This sampling technique helps provide yet another mechanism for trading off robustness for one threat model against another, as per design or specification requirements. For example, while both ablation models A7, A8 achieve similar ℓ_p-union robustness (42.4% and 42%), on the specific ℓ_1 and ℓ_∞ threat models, A7 achieves 48.8% and 44.7% robust accuracy respectively, while A8 achieves 45.1% and 46.5% robust accuracy respectively. This clearly indicates the trade-off achieved with sampling, wherein with $p = 0.6$, the model achieves higher ℓ_∞ robustness, alongside a reduction in ℓ_1 accuracy.

A7. Stability of NCAT

Table A3: **Stability across Reruns** Prediction accuracy (%) of ResNet-18 models trained on the ℓ_1 threat model using NCAT-ℓ_1 (left), and on the union of ℓ_1, ℓ_2 and ℓ_∞ threat models using NCAT (right). Robust accuracy is computed using only ℓ_1 adversaries in the left partition, while worst-case accuracy over adversaries constrained under the union of ℓ_1, ℓ_2 and ℓ_∞ threat models is presented in the right partition.

NCAT-ℓ_1	Clean Acc	ℓ_1 Robust Acc	NCAT	Clean Acc	Worst-Case Acc
Rerun-1	80.71	55.60	Rerun-1	80.46	42.58
Rerun-2	80.43	55.67	Rerun-2	80.52	42.51
Rerun-3	80.56	55.32	Rerun-3	80.38	42.45
Rerun-4	80.39	55.43	Rerun-4	80.56	42.27
Rerun-5	80.60	55.51	Rerun-5	80.47	42.46
Mean	80.54	55.51	Mean	80.48	42.45
Std-Dev	0.13	0.14	Std-Dev	0.07	0.11

In Table-A3, we analyze the variation in predictions accuracy for both clean and adversarial samples, for ResNet-18

models trained on CIFAR-10 using five different random seeds on the same Nvidia RTX 2080 TI GPU, with hyper-parameter frozen across reruns. In the left-partition of the table, we present results for the model trained to be robust against ℓ_1 adversaries in particular, using NCAT-ℓ_1, while in the partition on the right, we present results for the model trained to be robust against adversaries under the union of ℓ_1, ℓ_2 and ℓ_∞ threat models. We observe that models trained using either NCAT-ℓ_1 or NCAT are very stable across reruns, with variance levels similar to that reported from multi-step training approaches such as PGD-AT (Madry et al., 2018; Rice et al., 2020) and TRADES (Zhang et al., 2019). Furthermore, we note that NCAT based adversarial training does not suffer from catastrophic failure during any of the runs, in sharp contrast to that seen from RFGSM or NuAT based training, wherein catastrophic failures are observed in almost every training run.