2024 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 22–25, 2024, LONDON, UK

DEEP ADVERSARIAL DEFENSE AGAINST MULTILEVEL-ℓ_P ATTACKS

Ren Wang* *Yuxuan Li** *Alfred Hero*†

* Dept. ECE, Illinois Institute of Technology † Dept. EECS, University of Michigan

2024 IEEE 34th International Workshop on Machine Learning for Signal Processing (MLSP) | 979-8-3503-7225-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/MLSP58920.2024.10734748

ABSTRACT

Deep learning models have shown considerable vulnerability to adversarial attacks, particularly as attacker strategies become more sophisticated. While traditional adversarial training (AT) techniques offer some resilience, they often focus on defending against a single type of attack, e.g., the ℓ_∞-norm attack, which can fail for other types. This paper introduces a computationally efficient multilevel ℓ_p defense, called the Efficient Robust Mode Connectivity (EMRC) method, which aims to enhance a deep learning model's resilience against multiple ℓ_p-norm attacks. Similar to analytical continuation approaches used in continuous optimization, the method blends two p-specific adversarially optimal models, the ℓ_1- and ℓ_∞-norm AT solutions, to provide good adversarial robustness for a range of p. We present experiments demonstrating that our approach performs better on various attacks as compared to AT-ℓ_∞, E-AT, and MSD, for datasets/architectures including: CIFAR-10, CIFAR-100 / PreResNet110, WideResNet, ViT-Base.

Index Terms— adversarial training, robustness, ℓ_p norm perturbations, mode connectivity, model ensemble

1. INTRODUCTION

Deep learning models have revolutionized numerous fields, offering innovative solutions to complex problems [1, 2]. However, their vulnerability to adversarial attacks remains a significant concern [3, 4], undermining their practical utility and reliability. Specifically, these models are sensitive to slight, yet strategic, perturbations in their input data, which can mislead them into making incorrect predictions. While several methods aim to defend against such adversarial manipulations, most focus on enhancing the model's resilience against attacks based on a single type of perturbation metric, often measured by a ℓ_p norm [5, 6, 7, 8] for specific $p \in [1, \infty]$. This focus creates a defensive blind spot, leaving models vulnerable to other types of adversarial perturbations. On the other hand, recent studies that aim to achieve universal robustness across multiple ℓ_p norms either suffer from high computational costs or do not entirely solve the universal robustness problem: maintaining robustness against different types of perturbations concurrently [9, 10, 11, 12].

This paper addresses the shortcomings of current methods by proposing universally robust models capable of countering diverse types of ℓ_p-norm adversarial attacks. Previous studies have identified the mode connectivity property, which suggests that a path of high accuracy and low loss exists between two well-trained models in the parameter space [13, 14, 15]. Building on this concept and the theoretical evidence that affine classifiers can withstand multiple types of ℓ_p attacks if they are already resistant to ℓ_1 and ℓ_∞ perturbations [9], our work presents a novel approach: Efficient Robust Mode Connectivity (ERMC) combined with Model Ensemble which waeves ℓ_1 and ℓ_∞ robustness into the fabric of mode connectivity to derive a new training methodology. This amalgamation enables the identification of parameter paths that remain highly resistant to both ℓ_1 and ℓ_∞ perturbations, and therefore, multiple types of ℓ_p norm perturbations. We introduce an optimized fine-tuning technique with reduced computational complexity. Lastly, we employ a model ensemble strategy to select and aggregate models from this robust path, further improving robustness. Specifically, the algorithm works as follows. We first train one endpoint model optimized for ℓ_∞-norm adversarial training then retrain the model to be optimal relative to the ℓ_1-norm. Using these two endpoint models, and leveraging the mode connectivity property of deep neural networks (DNN), we identify a low-loss, high-robustness path connecting these endpoints. Finally, we deploy ensemble model aggregation to select models along this path that exhibit collective robustness against all types of ℓ_p-norm attacks, $1 \leq p \leq \infty$.

Contributions. We summarize our contributions below.

1. We improve upon traditional mode connectivity approaches to the design of DNN by integrating adversarial robustness, thereby uncovering a path that links an ℓ_∞ and an ℓ_1 adversarially trained model. This path demonstrates high resistance to other ℓ_p-norm attacks for $p \in [1, \infty]$.

2. We propose an Efficient Robust Mode Connectivity (ERMC) method, supplemented with model ensemble aggregation, that results in an efficient adversarial training algorithm with enhanced robustness.

Alfred Hero was supported in part by the US National Science Foundation under Grant CCF-2246213 and the US Army Research Office under Grant W911NF-23-1-0343. Ren Wang was supported by the US National Science Foundation under Grant 2246157 and the ORAU Ralph E. Powe Junior Faculty Enhancement Award.

3. Numerical experiments demonstrate that the proposed ERMC with model ensemble has superior performance in robustness against various ℓ_p attack modalities when compared to baseline approaches.

The rest of this article is structured as follows: Section 2 introduces related research on single-attack adversarial strategies and countermeasures, as well as defenses against a variety of ℓ_p norm perturbations. The subsequent Section 3 on multilevel ℓ_p-defense delves into the specifics of adversarial training and the optimization of robustness against multilevel ℓ_p perturbations. It lays the theoretical groundwork for our approach and addresses the question of achieving concurrent high robustness against both ℓ_1 and ℓ_∞ perturbations. Section 4 presents our novel ERMC approach in detail, describing how it incorporates robustness into mode connectivity and the ensemble model strategy used to boost robustness. Section 5 reports on the datasets, model architectures, evaluation methods, and the comprehensive experimental results, showcasing the effectiveness of ERMC. The Conclusion Section 6 summarizes our findings and contributions.

2. BACKGROUND AND RELATED WORK

2.1. Adversarial Attacks And Defenses

Recent studies have revealed that conventional machine learning models are susceptible to adversarially modified datasets. For a model $\boldsymbol{\theta}$ an adversary can target each feature $\mathbf{x} \in \mathbb{R}^d$ in a database $\mathcal{D}$ of feature-label pairs $\mathcal{D} = \{\mathbf{x}, y\}$, by solving the following *attacker's optimization* problem:

$$\arg\max_{\mathbf{x}'} \mathcal{L}(\boldsymbol{\theta}; \mathbf{x}', y), \quad s.t.\ d_p(\mathbf{x}', \mathbf{x}) \leq \epsilon_p. \tag{1}$$

Here $\mathcal{L}$ represents the training loss, e.g., the cross-entropy loss. ϵ_p is the attack-strength parameter of the type-p attacker, and d_p is a distance metric of type-p over the model parameter space. As in many other studies, we restrict attention to the case that d_p is the ℓ_p norm with $p \in [1, \infty]$. The solution to (1) is commonly known as the ℓ_p adversarial attack [5]. This problem is often iteratively solved using the fast gradient sign method [3] or projected gradient descent (PGD) [5], which computes the gradient $\nabla_{\mathbf{x}'}\mathcal{L}(\theta; \mathbf{x}', y)$ combined with a projection that constrains the perturbation $\mathbf{x}' - \mathbf{x}$ to the ℓ_p-ball of radius ϵ_p. The projection for the ℓ_p adversarial attack is denoted by $P_{\boldsymbol{\epsilon}_p}$. These methods may result in suboptimal solutions to (1) due to incorrect hyper-parameter tuning and gradient masking. To address these issues, methods such as the Auto Attack (AA) [16] and Multi Steepest Descent (MSD) [11] were introduced. Adversarial attacks can operate in a black-box manner, meaning the attacker does not have access to the model's parameters [17, 18]. However, this paper focuses only on scenarios where the attacker is aware of the model's parameters. To counter the attackers strategy (1), adversarial training (AT) methods are effective defense mechanisms [5, 6, 7, 8]. However, these methods often focus on a single type of ℓ_p disturbance, leading to decreased robustness against different types of perturbations [19].

2.2. Robustness Towards Multiple ℓ_p Norm Perturbations

In [10] the authors propose training on ℓ_∞-generated adversarial examples while selectively discarding inputs having low confidence scores, showing empirically that this results in a degree of robustness to ℓ_p-attacks for $p = 0, 1, 2, \infty$. The authors of [19] propose calculating the worst case attack by either picking the attack type that leads to the maximum loss or averaging the loss across all attack types. The Multi Steepest Descent (MSD) Defense [11] integrates multiple perturbation schemes to yield a more comprehensive ℓ_p robustness. The work in [9] offers a theoretically guaranteed defense mechanism but it only applies to affine classifiers. The Extreme Norm Adversarial Training (E-AT) method [20] employs a form of fine-tuning to practically implement the pathway from [9] and to reduce AT computational load. In contrast, in this paper we exploit the mode connectivity property of deep neural networks [13, 14, 15] to define the ERMC method that improves on the performance reported in [20].

3. MULTILEVEL ℓ_P-DEFENSE

Adversarial training (AT). Complementing the attacker's optimization (1), the defender aims to solve the *defender's optimization* problem:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{(\mathbf{x},y)\in\mathcal{D}} \left[\max_{\mathbf{x}':d_p(\mathbf{x}',\mathbf{x})\leq\epsilon_p} \mathcal{L}(\boldsymbol{\theta}; \mathbf{x}', y) \right], \tag{2}$$

using training data from $\mathcal{D}$ to empirically estimate the statistical expectation in (2), resulting in a solution we call Adversarial Training (AT)-ℓ_p. The main issue addressed in this section is that the solution AT-ℓ_p for a given p does not ensure robustness to other values of p in $[1, \infty]$. Furthermore, while in principle one could compute a dense set of solutions $\{\text{AT-}\ell_p\}_{p\in[1,\infty]}$, it is not clear how such solutions could be computed and combined in a computationally tractable manner to provide robustness over a range of p [19].

Optimizing robustness against multilevel ℓ_p perturbations. As argued in [9], affine and piecewise affine classifiers (like CNN with ReLU) can resist multiple ℓ_p norm attacks if they are already robust to ℓ_1 and ℓ_∞ perturbations. Specifically, Theorem 3.1 in [9] states that the convex hull of the union ball of the ℓ_1 and ℓ_∞ provides satisfactory robustness to ℓ_p perturbations, $1 \leq p \leq \infty$:

Theorem 1 *[9] Suppose that the classifier is piecewise affine. Let C be the convex hull of the union ball of the ℓ_1 and ℓ_∞. If $d \geq 2$ and $\epsilon_1 \in (\epsilon_\infty, d\epsilon_\infty)$, then*

$$\min_{\mathbb{R}^d\setminus C} \|\mathbf{x}' - \mathbf{x}\|_p = \frac{\boldsymbol{\epsilon}_1}{(\boldsymbol{\epsilon}_1/\boldsymbol{\epsilon}_\infty - \beta + \beta^q)^{1/q}} \tag{3}$$

where $\beta = \frac{\epsilon_1}{\epsilon_\infty} - \lfloor \frac{\epsilon_1}{\epsilon_\infty} \rfloor$ *and* $\frac{1}{p} + \frac{1}{q} = 1$.

The salient question arising is: how can one concurrently achieve high robustness against both ℓ_1 and ℓ_∞ perturbations? To address this question a cutting-edge study, E-AT [20], proposed using a method called fine-tuning to efficiently update the model from AT-ℓ_∞ to AT-ℓ_1, asserting that this results in robustness to a range of ℓ_p disturbances. Yet, two notable issues persist: ❶ while the fine-tuned model may exhibit robustness against ℓ_1-norm attacks, it may have lost some robustness against the original ℓ_∞-norm attack; and ❷ Achieving both high ℓ_∞ robustness and high ℓ_1 robustness is inherently challenging for a single model, given its limited capacity. To address these dual challenges, this paper introduces a mode-connectivity-based approach that simultaneously identifies a large number of models having both high ℓ_∞ robustness and high ℓ_1 robustness. This results in a larger union ball, thereby enhancing the model's resilience against a broader range of perturbations.

4. PROPOSED METHODS

We aim to improve the joint robustness to both ℓ_∞ and ℓ_1 perturbations by leveraging two adversarially trained models.

4.1. Incorporating robustness into mode connectivity

For neural networks, mode connectivity is the property that pairs of local minima (modes) discovered by gradient-based optimization techniques are connected through simple paths over which the model's loss does not change appreciably [13, 14]. In [14] mode connectivity is established for a wide range of DNNs and datasets. The path between a pair of modes $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ is constructed over the parameter space of the neural network by minimizing the averaged loss function, $\mathcal{L}$, over all possible simple paths. The path is represented as $\phi_{\boldsymbol{\theta}} = \{\phi_{\boldsymbol{\theta}}(t), t \in [0,1]\}$, where $\boldsymbol{\theta}$ is a free parameter, which satisfies the endpoint conditions $\phi_{\boldsymbol{\theta}}(0) = \boldsymbol{\theta}_1$ and $\phi_{\boldsymbol{\theta}}(1) = \boldsymbol{\theta}_2$. Specifically, to find a desired low-loss path between the modes $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, one minimizes the following statistical expectation

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{t\sim U(0,1)} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \mathcal{L}(\phi_{\boldsymbol{\theta}}(t); (\mathbf{x}, y)), \tag{4}$$

where $U(0,1)$ represents the uniform distribution over the interval $[0,1]$. The curve $\phi_{\boldsymbol{\theta}}$ is fixed as a Quadratic Bezier Curve (QBC) [21] across this paper:

$$\phi_{\boldsymbol{\theta}}(t) = (1-t)^2\boldsymbol{\theta}_1 + 2t(1-t)\boldsymbol{\theta} + t^2\boldsymbol{\theta}_2. \tag{5}$$

The main assumption behind this paper is that the notion of mode connectivity can be extended to adversarial loss functions associated with different ℓ_p-types, resulting in paths that maintain a high level of robustness against ℓ_∞ and ℓ_1 attacks, in addition to improving robustness to other ℓ_p attacks. The proposed extension consists of two additional steps: **(Step 1)** The endpoint parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are trained via AT-ℓ_∞ and AT-ℓ_1; **(Step 2)** We solve the following modification of (4) to preserve adversarial robustness for $p \in \{1, \infty\}$:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{t\sim U(0,1)} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \sum_{p\in\{1,\infty\}} \max_{d_p(\mathbf{x}',\mathbf{x})\le\epsilon_p} \mathcal{L}(\phi_{\boldsymbol{\theta}}(t); (\mathbf{x}', y)), \tag{6}$$

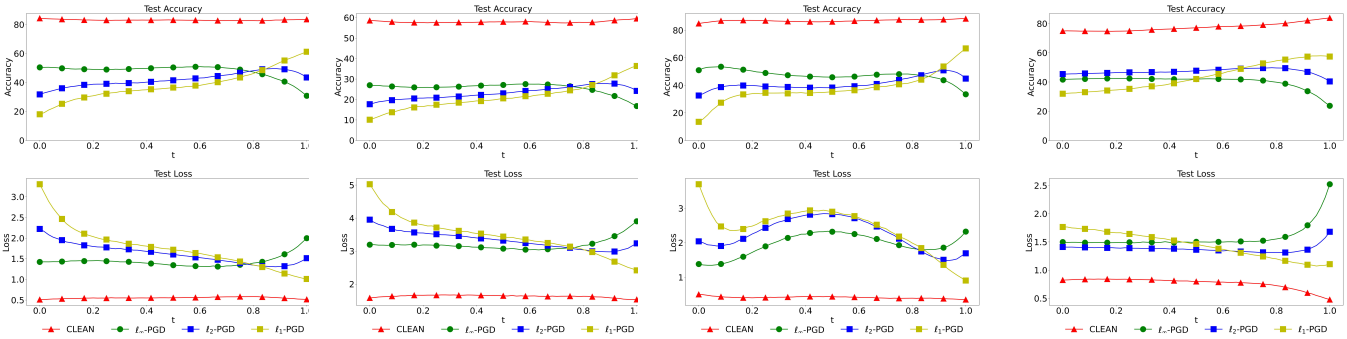
where $\phi_{\boldsymbol{\theta}}(0)$ and $\phi_{\boldsymbol{\theta}}(1)$ are the two AT models, AT-ℓ_∞ and AT-ℓ_1, respectively. In the inner optimization loop d_p corresponds to the ℓ_∞ and ℓ_1 distances for $p = 0$ and $p = 1$. We use a Multi Steepest Descent (MSD) technique to solve the maximization in the inner loop that encompasses both ℓ_∞ and ℓ_1 perturbations within each step of PGD, similarly to [11]. In each epoch, for every data batch, we randomly choose a value for t. The subsequent training closely resembles Adversarial Training (AT), with the key difference being that we pick the worst-case perturbation from two types of perturbations in each inner loop iteration. Consequently, the algorithmic complexity remains similar to that of standard AT.

Algorithm 1 Efficient Robust Mode Connectivity

Require: A model $\phi_{\boldsymbol{\theta}}(0)$ trained with AT-ℓ_∞; initial model $\boldsymbol{\theta}^0$; the corresponding projections $P_{\boldsymbol{\delta}_1}$ and $P_{\boldsymbol{\delta}_\infty}$; training set $\mathcal{D}$; iteration number J; batch size B; initial perturbation $\boldsymbol{\delta}^{(0)} = \mathbf{0}$.

Create a copy of $\phi_{\boldsymbol{\theta}}(0)$ and retrain it with AT-ℓ_1 for 10 epochs to obtain a model $\phi_{\boldsymbol{\theta}}(1)$.
$\boldsymbol{\theta} = \boldsymbol{\theta}^0$
for each data batch $\mathcal{D}_b \in \mathcal{D}$ in each epoch $e \in E$ **do**
  Uniformly select $t \sim U(0,1)$
  for $\forall \mathbf{x} \in \mathcal{D}_\mathbf{b}$ **do**
    for $j = 1, \cdots, J$ **do**
      $\boldsymbol{\delta}_1^{(j)} \leftarrow P_{\boldsymbol{\epsilon}_1}\big(\boldsymbol{\delta}^{(j-1)} - \nabla_{\boldsymbol{\delta}}\mathcal{L}(\phi_{\boldsymbol{\theta}}(t); \mathbf{x} + \boldsymbol{\delta}^{(j-1)}, y)\big)$
      $\boldsymbol{\delta}_\infty^{(j)} \leftarrow P_{\boldsymbol{\epsilon}_\infty}\big(\boldsymbol{\delta}^{(j-1)} - \nabla_{\boldsymbol{\delta}}\mathcal{L}(\phi_{\boldsymbol{\theta}}(t); \mathbf{x} + \boldsymbol{\delta}^{(j-1)}, y)\big)$
    end for
    $\boldsymbol{\delta}^{(j)} \leftarrow \arg\max_{\boldsymbol{\delta}_i^{(j)}, i\in\{1,\infty\}} \mathcal{L}(\phi_{\boldsymbol{\theta}}(t); \mathbf{x} + \boldsymbol{\delta}_i^{(j)}, y)$
  end for
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \nabla_{\boldsymbol{\theta}} \sum_{\mathbf{x}\in\mathcal{D}_\mathbf{b}} \mathcal{L}(\phi(t;\boldsymbol{\theta}); \mathbf{x} + \boldsymbol{\delta}^{(j)}, y)$
end for
return $\boldsymbol{\theta}, \phi_{\boldsymbol{\theta}}(t), \forall t \in [0,1]$

We conclude this sub-section by noting that the concept of expansion of the set of high adversarially robust models beyond two models AT-ℓ_1 and AT-ℓ_∞ is similar to the concept of analytic continuation in complex analysis, more specifically the converse analytic continuation method called blending, which seeks to extend two analytic functions defined over disjoint domains to a single C^∞ function over a path connecting the domains [22].



(a) CIFAR-10 (b) CIFAR-100 (c) WideResNet-28-10 (d) ViT-base

Fig. 1: ERMC can find paths with high robustness against $\ell_\infty/\ell_2/\ell_1$ attacks by connecting a ℓ_∞ model and a ℓ_1 model. The effectiveness of ERMC is validated on different datasets and model architectures. Upper panels: the accuracy of the clean test and the robust accuracies under $\ell_\infty/\ell_2/\ell_1$-PGD attacks. Lower panels: the associated loss values of clean test data and perturbed test data. (a) and (b): results obtained from the CIFAR-10 and CIFAR-100 datasets, using the PreResNet110 model architecture. (c) and (d): results from the CIFAR-10 dataset, utilizing the WideResNet-28-10 and ViT-base model architectures.

4.2. ERMC with model ensemble

We reduce the computation burden of solving two independent AT-ℓ_p problems, for $p = \infty$ and $p = 1$, by introducing a more efficient approach: the efficient robust model connectivity algorithm. In ERMC, initially a model with high robustness to either ℓ_∞ or ℓ_1 perturbation is trained, after which a copy is created and retrained for another few epochs under the other perturbation type using the same training set. Similarly to its use in E-AT [20], the fine-tuning step provides more efficient computation of the AT-ℓ_∞ and AT-ℓ_1 adversarial models in ERMC. In the experiments described below, the number of fine-tuning epochs is set to 10 yielding a computationally less burdensome determination of the second endpoint, while retaining the first one, facilitating the identification of a high-robustness path as provided by (6). The full algorithm of ERMC is presented in Algorithm 1. The second endpoint $\phi_{\boldsymbol{\theta}}(1)$ is trained from the first endpoint $\phi_{\boldsymbol{\theta}}(0)$ using a different perturbation type. In each epoch, we sample a t uniformly from the uniform distribution. Then, in each iteration of generating perturbations, we consider two types: ℓ_∞ and ℓ_1. Subsequently, we select the perturbations that cause the highest losses and use them to update the model parameters. The number of iterations, denoted by J, is set at 10.

We have observed in experiments that certain regions along the path contain models that exhibit high levels of robustness for both types of perturbations. The optimal model along the path can be identified by assessing the trajectory with lower robust accuracy under ℓ_∞ and ℓ_1 attacks, and then selecting the point that performs best in this worst-case scenario. This single-model approach offers the advantage of circumventing the limitations inherent to E-AT [20] while capitalizing on robustness against both types of perturbations. However, given the existence of many models along the path that exhibit high degrees of robustness to ℓ_∞ and ℓ_1 attacks, it's natural to consider a model ensemble strategy to further bolster performance. This leads to a model that is collectively more robust to both ℓ_∞ and ℓ_1 perturbations. The ensemble selection proceeds as follows. We find a segment $[a, b]$ along the path $\phi_{\boldsymbol{\theta}}$ satisfying the criterion: each point on the segment has robust accuracies surpassing two prefixed *model selection thresholds* α_∞, α_1 under ℓ_∞ and ℓ_1 attacks, respectively. We then choose $n > 1$ models at path locations given by $t = a + \frac{b-a}{n-1}i$, where i ranges from 0 to $n-1$. If multiple non-continuous intervals meet the above criterion, the n points can be distributed among them proportionately to their respective lengths. We denote ERMC with n selected models as ERMC-n and average the outputs of these n models' final layers to form our class probability prediction.

5. EXPERIMENTS

Dataset selection and model architectures. We test our proposed techniques on CIFAR-10 (as the default dataset) and CIFAR-100 [23] datasets, utilizing PreResNet110 (as the default architecture), WideResNet-28-10, and Vision Transformer-base (ViT-base).

Evaluation methods and metrics. We set the attack strength parameters constraining the ℓ_∞, ℓ_2, and ℓ_1 norms to the commonly used values $\epsilon = 8/255, 1$, and 12, respectively. In our evaluation, we implemented basic PGD adversarial attacks as well as Auto-Attack (AA) [16] under $\ell_\infty, \ell_2, \ell_1$ norm perturbations, in addition to implementing the MSD attack. Metrics for assessment include: ① Standard accuracy (SA) on clean test data; ② Robust accuracies under various perturbation types including $\ell_\infty/\ell_2/\ell_1$-PGD, MSD attack, and $\ell_\infty/\ell_2/\ell_1$ AA; and ③ Sample-wise worst-case scenario accu-

racy (Union) calculated from all three basic PGD adversarial methods. A sample is considered correct only if it is accurately predicted under each of the three basic PGD adversarial attacks. These experiments were run on two NVIDIA RTX A100 GPUs.

Table 1: Our Method Achieves State-Of-The-art Robustness Levels Under Various Perturbations on CIFAR-10. ERMC surpasses baseline performance without the use of an ensemble. The best results are in **bold**.

	SA	PGD ($\ell_\infty/\ell_2/\ell_1$)	Union	AA [16] ($\ell_\infty/\ell_2/\ell_1$)	MSD
AT-ℓ_∞ [5]	85.00%	49.03%/29.66%/{16.61%}	21.85%	46.02%/20.86%/{10.45%}	15.27%
MSD [11] Defense	81.35%	{40.14%}/48.58%/47.50%	38.35%	{37.87%}/45.9%/45.27%	38.20%
E-AT [20]	79.3%	{44.07%}/49.12%/49.82%	41.08%	{41.41%}/46.5%/47.82%	42.67%
ERMC-1 (ours $n = 1$)	82.66%	{46.54%}/48.76%/47.06%	41.94%	44.88%/45.88%/{43.97%}	44.88%
ERMC-3 (ours $n = 3$)	79.61%	49.29%/51.32%/{48.49%}	45.27%	{42.88%}/44.57%/47.37%	43.31%
ERMC-5 (ours $n = 5$)	79.41%	55.46%/57.28%/{**53.97%**}	**51.41%**	{**49.33%**}/50.55%/52.41%	**49.78%**

Experimental results. As a baseline, endpoint models are trained for 150 epochs, with paths derived through an extra 50 epochs. The models at the left (right) endpoints are trained with AT-ℓ_∞ (AT-ℓ_∞ and fine-tuned with AT-ℓ_1). The results are displayed in Fig. 1. The upper panels show the clean test accuracy and accuracies under $\ell_\infty/\ell_2/\ell_1$-PGD attacks. The lower panels show the corresponding loss values. t varies from 0 to 1. Moving from left to right in Fig. 1, panels (a) and (b) depict results obtained from the CIFAR-10 and CIFAR-100 datasets, respectively, using the PreResNet110 model architecture. Conversely, panels (c) and (d) present results from the CIFAR-10 dataset, but utilizing the WideResNet-28-10 and ViT-base model architectures. Notably, we find: ❶ The existence of robust paths, which shows that the ERMC application enhances resilience to multiple attack types, although they don't form straight lines like in mode connectivity; ❷ ERMC performs well on all considered datasets and architectures; ❸ The robust paths also function as effective mode connectivity paths, where both the clean accuracy and loss (indicated by red lines) maintain consistent levels between the two endpoints $t = 0$ and $t = 1$; and ❹ Fine-tuning influences original robustness levels, where endpoint models show strong resilience against corresponding perturbation types but are weaker against others. For example, the left (right) endpoint has a high resilience to ℓ_∞ (ℓ_1) perturbations but suffers from attacks using ℓ_1 (ℓ_∞) perturbations. Additionally, ERMC reduces the required computation time by approximately 36% on a single GPU relative to the brute force approach of solving AT-ℓ_∞ and AT-ℓ_1 independently.

Comparative analyses with different baselines are summarized in Table 1. We evaluate them using all aforementioned metrics, and the lowest accuracy under the three basic ℓ_p-PGD attacks (and three ℓ_p AA) are indicated within braces. These baselines - comprising AT-ℓ_∞ [5], E-AT [20], and MSD Defense [11] - are trained over 200 epochs. The model selection thresholds are set at $\alpha_\infty = 37\%$ for ℓ_∞ robustness and $\alpha_1 = 43\%$ for ℓ_1 robustness. As per Table 1, observe that ERMC-1 outperforms MSD Defense (and E-AT) in terms of accuracy improvements under various metrics, indicated by percentages 6.4%, 3.59%, 6.1%, and 6.68% (2.47%, 0.86%, 2.56%, and 2.21%) under $\ell_\infty/\ell_2/\ell_1$-PGD, Union, $\ell_\infty/\ell_2/\ell_1$ AA, and MSD Attack, respectively. It is also observable that as the number of models n increases, the performance of ERMC correspondingly improves. When n reaches to 5, ERMC-5 outperforms MSD Defense (and E-AT) in terms of accuracy improvements under various metrics, indicated by percentages 13.85%, 13.06%, 11.46%, and 11.58% (9.9%, 10.33%, 7.92%, and 7.11%) under $\ell_\infty/\ell_2/\ell_1$-PGD, Union, $\ell_\infty/\ell_2/\ell_1$ AA, and MSD Attack, respectively. It's crucial to highlight that our method surpasses baseline performance without the use of an ensemble. The further enhancement observed with an ensemble simply underscores the value added by ensemble boosting of ERMC's baseline performance from the single model context. Unlike baselines that require multiple runs to generate a similar number of models, our approach naturally produces a model population in a single run, offering an attractive time-efficient alternative. The ERMC approach demonstrates a trade-off between clean accuracy and robustness. Nonetheless, the decrease in clean accuracy, quantified at 2.34% when measured against AT-ℓ_∞, is more modest compared to the degradation suffered by other defensive strategies like MSD Defense and E-AT.

6. CONCLUSION

This paper introduces the Efficient Robust Mode Connectivity (ERMC) method, a novel approach for enhancing the resilience of deep learning models against various adversarial ℓ_p-norm attacks. By combining the robustness benefits of ℓ_1 and ℓ_∞ adversarial training within a single framework, ERMC transcends the limitations of traditional methods that focus on single-type perturbations. Leveraging mode connectivity theory with efficient tuning and ensemble strategies, the method achieves a robust defense. Experimental results show that ERMC outperforms established defenses like AT-ℓ_∞, E-AT, and MSD Defense, particularly against ℓ_∞ and ℓ_1 perturbations and other ℓ_p-norm attacks. Its integration of multiple adversarial training types enhances defense capabilities while preserving efficiency, marking a significant step forward in adversarial robustness and suggesting new directions for further research in the security of deep learning.

7. REFERENCES

[1] Wenting Li, Deepjyoti Deka, Ren Wang, and Mario R Arrieta Paternina, "Physics-constrained adversarial training for neural networks in stochastic power grids," *IEEE Transactions on Artificial Intelligence*, 2023.

[2] Melanie Jouaiti and Kerstin Dautenhahn, "Dysfluency classification in stuttered speech using deep learning for

real-time applications," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6482–6486.

[3] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[4] Lichao Sun, Yingtong Dou, Carl Yang, Kai Zhang, Ji Wang, S Yu Philip, Lifang He, and Bo Li, "Adversarial attack and defense on graph data: A survey," *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[5] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.

[6] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein, "Adversarial training for free!," in *Advances in Neural Information Processing Systems*, 2019, pp. 3353–3364.

[7] Ren Wang, Kaidi Xu, Sijia Liu, Pin-Yu Chen, Tsui-Wei Weng, Chuang Gan, and Meng Wang, "On fast adversarial robustness adaptation in model-agnostic meta-learning," in *International Conference on Learning Representations*, 2020.

[8] Eric Wong, Leslie Rice, and J. Zico Kolter, "Fast is better than free: Revisiting adversarial training," in *International Conference on Learning Representations*, 2020.

[9] Francesco Croce and Matthias Hein, "Provable robustness against all adversarial ℓ_p-perturbations for $p \geq 1$," in *International Conference on Learning Representations*, 2020.

[10] David Stutz, Matthias Hein, and Bernt Schiele, "Confidence-calibrated adversarial training: Generalizing to unseen attacks," in *International Conference on Machine Learning*. PMLR, 2020, pp. 9155–9166.

[11] Pratyush Maini, Eric Wong, and Zico Kolter, "Adversarial robustness against the union of multiple perturbation models," in *International Conference on Machine Learning*. PMLR, 2020, pp. 6640–6650.

[12] Ren Wang, Tianqi Chen, Stephen M Lindsly, Cooper M Stansbury, Alnawaz Rehemtulla, Indika Rajapakse, and Alfred O Hero, "Rails: A robust adversarial immune-inspired learning system," *IEEE Access*, vol. 10, pp. 22061–22078, 2022.

[13] C Daniel Freeman and Joan Bruna, "Topology and geometry of half-rectified network optimization," in *Int. Conf. on Learning Representation (ICLR)*, 2017.

[14] Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson, "Loss surfaces, mode connectivity, and fast ensembling of dnns," *Advances in neural information processing systems*, vol. 31, 2018.

[15] Ren Wang, Yuxuan Li, and Sijia Liu, "Exploring diversified adversarial robustness in neural networks via robust mode connectivity," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2345–2351.

[16] Francesco Croce and Matthias Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *International Conference on Machine Learning*. PMLR, 2020, pp. 2206–2216.

[17] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein, "Square attack: a query-efficient black-box adversarial attack via random search," in *European conference on computer vision*. Springer, 2020, pp. 484–501.

[18] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh, "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*. ACM, 2017, pp. 15–26.

[19] Florian Tramer and Dan Boneh, "Adversarial training and robustness for multiple perturbations," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[20] Francesco Croce and Matthias Hein, "Adversarial robustness against multiple and single ℓ_p-threat models via quick fine-tuning of robust classifiers," in *International Conference on Machine Learning*. PMLR, 2022, pp. 4436–4454.

[21] Rida T Farouki, "The bernstein polynomial basis: A centennial retrospective," *Computer Aided Geometric Design*, vol. 29, no. 6, pp. 379–419, 2012.

[22] Lloyd N Trefethen, "Numerical analytic continuation," *Japan Journal of Industrial and Applied Mathematics*, pp. 1–50, 2023.

[23] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," *Master's thesis, Department of Computer Science, University of Toronto*, 2009.