Exploring Diversified Adversarial Robustness in Neural Networks via Robust Mode Connectivity

Ren Wang[1]*, Yuxuan Li[1,2]†, Sijia Liu[3]
[1]Illinois Institute of Technology, Chicago, IL, US
[2]Harbin Institute of Technology, Harbin, Heilongjiang, China
[3]Michigan State University, East Lansing, MI, US

rwang74@iit.edu lyxzcx@outlook.com liusiji5@msu.edu

Abstract

This paper proposes a new method called robust mode connectivity (RMC) to enhance the adversarial robustness of neural networks (NNs) by exploring a wider range of parameter space. While adversarial training methods have shown promising results in enhancing the robustness of NNs against perturbations, they are limited by considering only a single type of perturbation during training and having limited search capability. RMC aims to address this limitation by considering multiple ℓ_p norm perturbations ($p = 1, 2, \infty$) and building on the concept of mode connectivity to identify a path of NNs with high robustness against different types of perturbations. The proposed method employs a multi steepest descent (MSD) algorithm to explore the parameter space and achieve diversified adversarial robustness. Experimental results on various datasets and architectures demonstrate the effectiveness of RMC.

1. Introduction

Over the past ten years, neural networks (NNs) have been widely utilized in various fields, such as healthcare [19], face recognition [6, 13], and power systems [1, 16] that require high security. NNs are essential components of deep learning, as they can learn the desired mappings from a given set of data. However, despite NNs' ability to accurately identify the underlying relationships in the data, they are sensitive to even the slightest changes in the inputs, known as adversarial perturbations [2, 11, 17, 26]. This vulnerability raises concerns about the trustworthiness of these models.

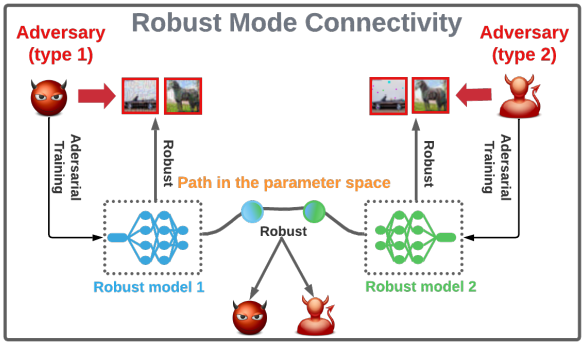


Figure 1. The Robust Mode Connectivity (RMC) method works by finding a path of neural network models that exhibit robustness to different types of adversarial attacks. Specifically, RMC seeks to connect two models that are robust to different types of attacks, such as adversary type 1 and adversary type 2. This path in the parameter space ensures enhanced robustness against adversary type 1 and adversary type 2.

Recent extensive research has focused on addressing the vulnerability issues of neural networks (NNs). Among these efforts, adversarial training and its variants have demonstrated outstanding performance [17, 20, 27]. Adversarial training aims to update NNs by continuously generating adversarial examples from clean training data, which helps NNs learn adversarial distributions and maintain a certain level of robustness during the inference phase. However, existing works mostly consider a single type of ℓ_p norm perturbation during adversarial training, resulting in a rapid decline in robustness when faced with perturbations different from those used in training [22]. Although some studies have attempted to address this issue by training NNs with multiple ℓ_p norm perturbations [3, 4, 18, 21–23],

*Corresponding author: Ren Wang.
†Work done during an internship at the Trustworthy and Intelligent Machine Learning Research Lab in the Department of Electrical and Computer Engineering, Illinois Institute of Technology.
This work was supported by the National Science Foundation (NSF) under Grant 2246157.

they have not completely resolved the lack of multiple adversarial robustness. Traditional neural network learning mechanisms rely on optimization of a single set of parameters. We hypothesize that the weakness stems from the narrow search scope of current approaches and suggest that a successful resilient learning technique must thoroughly explore a wider range. One technique that meets the requirement is population-based optimization, which can explore a broader range of solutions by maintaining a diverse population of candidate solutions and optimize complex problems. Such methods typically neglect adversarial robustness [5], have low learning speed [14], or only work in the input space [24,25]. Recent studies have found that low-loss high-accuracy paths exist in the parameter space, which is named mode connectivity and the paths can be found using an accelerated population-based optimization strategy [9]. Nonetheless, employing this technique alone proves inadequate in the adversarial scenario.

Based on the mode connectivity property and the hypothesis that conventional training methods lack space search, we propose a new method called robust mode connectivity (RMC). The primary objective of RMC is to find paths that enhance the diversified ℓ_p robustness of models when faced with attacks that adhere to ℓ_p norms. In this paper, we consider $p = 1, 2, \infty$. The RMC method builds on the idea of mode connectivity and employs a multi steepest descent (MSD) algorithm [18] to identify a path of NNs that exhibit high robustness against different types of perturbations, as depicted in Figure 1. RMC establishes a path in the parameter space that connects two neural network models, each with robustness to different types of adversarial attacks. This path contain points that enhance the overall robustness against both types of attacks, providing a more effective defense mechanism.

The remaining sections of this paper are structured as follows. In Section 2, we begin with a pilot study on injecting robustness into vanilla mode connectivity. Section 3 presents the proposed RMC method in detail. In Section 4, we report on the experimental results that validate the effectiveness of the RMC method. Finally, we summarize our findings and discuss future research directions in Section 5.

2. Exploring the Attainment of Robust Path with Vanilla Mode Connectivity: A Pilot Study

2.1. Mode Connectivity

Traditional training methods usually train a model from scratch, searching the parameter space starting from a random initial point. Such training methods, in most cases, will converge to a local minimum. Mode connectivity is a property of neural networks where simple paths between local minimums found by gradient descent methods exist in the parameter space [8, 9]. The cost function along the path is similar to the endpoints, which are two sets of neural network parameters $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d$ trained by minimizing a given loss $\mathcal{L}$. The parameter curve $\phi(t;\boldsymbol{\theta}) \in \mathbb{R}^d, t \in [0,1]$ is a smooth representation of the path, where $\phi(0;\boldsymbol{\theta}) = \boldsymbol{\theta}_1$ and $\phi(1;\boldsymbol{\theta}) = \boldsymbol{\theta}_2$.

To find a low-loss path between $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, we minimize the following expectation over a uniform distribution on the curve:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{t\sim q(t;\boldsymbol{\theta})} \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}_0} \mathcal{L}(\phi(t;\boldsymbol{\theta});(\mathbf{x},\mathbf{y})), \tag{1}$$

where $\mathcal{D}_0$ denotes the benign dataset. $q(t;\boldsymbol{\theta})$ represents the distribution for sampling the parameters on the path. Note that using stochastic gradient descent on (1) is generally intractable. Therefore a computationally tractable surrogate is proposed as follows.

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{t\sim U(0,1)} \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}_0} \mathcal{L}(\phi(t;\boldsymbol{\theta});(\mathbf{x},\mathbf{y})), \tag{2}$$

where $U(0,1)$ denotes the uniform distribution on $[0,1]$. Bezier curves [7] and Polygonal chains [10] are commonly used in mode connectivity to serve as $\phi(t;\boldsymbol{\theta})$. Training neural networks on these curves provides many similar-performing models on low-loss paths. Trough out this paper, we will use quadratic Bezier curve, which is defined as $\phi(t;\boldsymbol{\theta}) = (1-t)^2\boldsymbol{\theta}_1 + 2t(1-t)\boldsymbol{\theta} + t^2\boldsymbol{\theta}_2$. Figure 2 shows a quadratic Bezier curve obtained from (2) that connects two models with near-constant loss. It can be seen that mode connectivity provides an efficient strategy to search the parameter space.

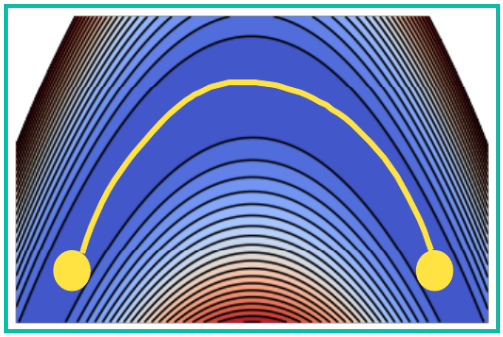


Figure 2. Mode connectivity in the parameter space found a path that exhibits nearly constant loss. The endpoints are two pretrained models. Quadratic Bezier curve is used in the search.

2.2. Vanilla Mode Connectivity on Robust Endpoints

It is evident that vanilla mode connectivity does not account for robustness nor does it address various types of adversarial attacks. The application of vanilla mode connectivity alone leads to models that lack robustness. However,

by configuring $\phi(0;\boldsymbol{\theta})$ and $\phi(1;\boldsymbol{\theta})$ as adversarially-trained neural networks that have been exposed to different types of perturbations, the application of equation (2) could potentially yield a path comprising points that exhibit a high degree of robustness against all types of perturbations.

$\phi(0;\boldsymbol{\theta})$ and $\phi(1;\boldsymbol{\theta})$ are firstly trained by the min-max optimization-based adversarial training (AT), as documented in [17]. AT can be summarized as follows:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{(\mathbf{x},y)\in\mathcal{D}_0}[\max_{\mathrm{Dist}_i(\mathbf{x}',\mathbf{x})\leq\delta_i} \mathcal{L}(\boldsymbol{\theta};\mathbf{x}',y)], \tag{3}$$

where the δ_is are sufficiently small values, and Dist_is are distance measurement functions. The inner maximization optimization is usually referred to as an adversarial attack [17]. In this paper, we restrict the distance measures $\mathrm{Dist}i$s to be ℓ_p norms, where $p = 1, 2, \mathrm{or}, \infty$. A practical way to solve the inner maximization optimization in (3) is to apply gradient descent and projection P_{δ_i} that maps the perturbation $\boldsymbol{\epsilon_i} = \mathbf{x}' - \mathbf{x}$ to a feasible set, which is usually referred to as the PGD attack. We will use ℓ_p-PGD to denote the PGD attack with the ℓ_p norm. We shall use the term ℓ_p-AT to refer to AT with the ℓ_p norm. In our setting, $\phi(0;\boldsymbol{\theta})$ and $\phi(1;\boldsymbol{\theta})$ are trained with two different ℓ_p-AT. Below we provide the detailed settings.

Settings of vanilla mode connectivity on robust endpoints. We combine two PreResNet110 models [12], one trained with ℓ_∞-AT ($\delta = 8/255$, 150 epochs) and the other trained with ℓ_2-AT ($\delta = 1$, 150 epochs), to find the desired path using the vanilla mode connectivity (2). The mode connectivity curve is generated by an additional 50 epochs of training.

The outcome is presented in Figure 3. The left (right) endpoint corresponds to the model trained with ℓ_∞-AT (ℓ_2-AT). It is evident that the path has a high loss and low robust accuracy on both types of attacks, although the accuracy on clean data is still high. This suggests that the vanilla mode connectivity approach is unable to locate a path that offers high robustness against ℓ_∞-PGD and ℓ_2-PGD attacks. The reason is that the search in the parameter space is still based on clean training data $\mathcal{D}_0$.

3. Robust Mode Connectivity

3.1. Mode Connectivity with Robust Search

While the vanilla mode connectivity seeks to reveal the underlying geometry of the loss landscape, it explores the search space based on the original data distribution. Hence, it cannot yield high robustness by solely utilizing two adversarially-trained models as endpoints. To address this issue, we establish a link between mode connectivity (2) and adversarial training under diversified ℓ_p adversarial perturbations. We modify the objective (2) to align with our goal. We introduce an adversarial generator in the inner maximization loop and incorporate various perturbation types in the generator to avoid robustness bias resulting from a single type of perturbation. As a result, we obtain

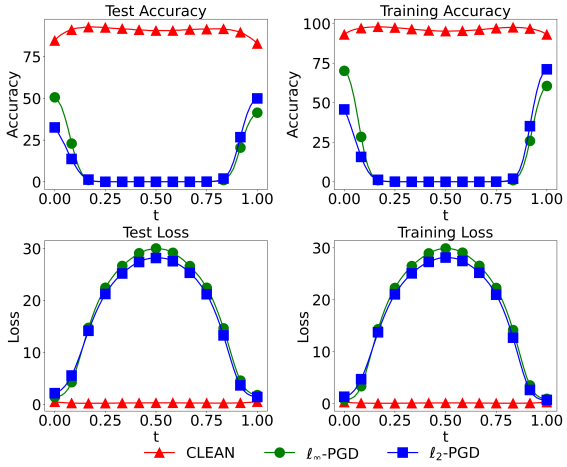
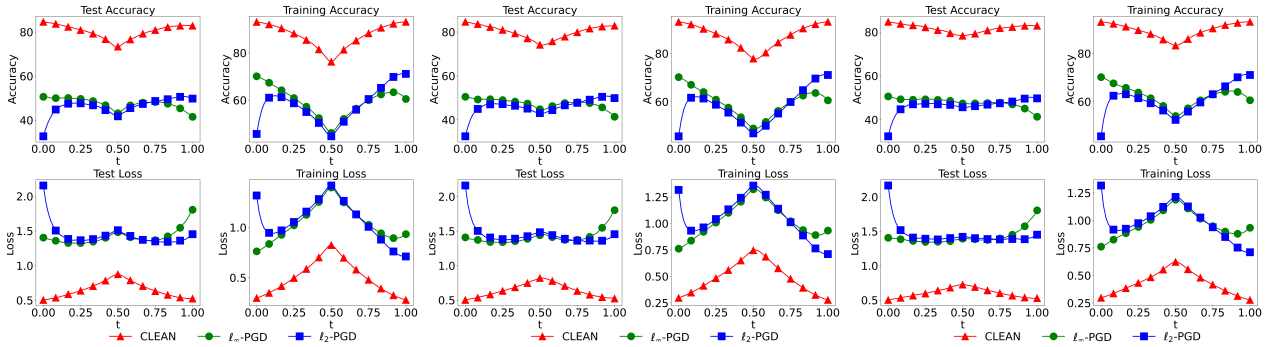


Figure 3. The vanilla mode connectivity (2) is unable to identify highly robust paths, despite using models trained with ℓ_∞-AT and ℓ_2-AT as two extreme points. The models $\phi(0;\boldsymbol{\theta})$ and $\phi(1;\boldsymbol{\theta})$ are trained with ℓ_∞-AT ($\delta = 8/255$, 150 epochs) and ℓ_2-AT ($\delta = 1$, 150 epochs), respectively.

Algorithm 1 Robust Mode Connectivity

Require: $\phi(0;\boldsymbol{\theta})$, $\phi(1;\boldsymbol{\theta})$ - two selected models with the same structure (potentially trained with different strategies, e.g., AT under different perturbation types); initial model $\boldsymbol{\theta}^0$; the perturbation types $i \in I$ and the corresponding projections P_{δ_i}; training set $\mathcal{D}_0$; inner loop iteration number J; batch size B; initial perturbation $\epsilon^{(0)} = \mathbf{0}$.

$\boldsymbol{\theta} = \boldsymbol{\theta}^0$.
For each data batch $\mathcal{D}_b \in \mathcal{D}_0$ 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**
    for $i \in I$ **do**
      $\boldsymbol{\epsilon}_i^{(j)} \leftarrow P_{\delta_i}[\boldsymbol{\epsilon}^{(j-1)} - \nabla_{\boldsymbol{\epsilon}}\mathcal{L}(\phi(t;\boldsymbol{\theta});\mathbf{x}+\boldsymbol{\epsilon}^{(\mathbf{j-1})},\mathbf{y})]$.
    end for
    $\boldsymbol{\epsilon}^{(j)} \leftarrow \arg\max_{\boldsymbol{\epsilon}_i^{(j)}, i\in I} \mathcal{L}(\phi(t;\boldsymbol{\theta});\mathbf{x}+\boldsymbol{\epsilon}_\mathbf{i}^{(\mathbf{j})},\mathbf{y})$.
  end for
  end For
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \nabla_{\boldsymbol{\theta}} \sum_{\mathbf{x}\in\mathcal{D}_\mathbf{b}} \mathcal{L}(\phi(t;\boldsymbol{\theta});\mathbf{x}+\boldsymbol{\epsilon}^{(\mathbf{j-1})},\mathbf{y})$
end For
return $\boldsymbol{\theta}$, $\phi(t;\boldsymbol{\theta}), \forall t \in [0,1]$



(a) Epoch=50 (b) Epoch=100 (c) Epoch=150

Figure 4. By utilizing models trained with ℓ_∞-AT and ℓ_2-AT as two endpoints, the RMC (4) can identify a highly robust path. The inner solver for solving (4) is MSD [18], which utilizes perturbations generated by ℓ_2 and ℓ_∞ norm distance measures. Panel (a)/(b)/(c) requires 50/100/150 epochs to solve.

a model path $\phi(t;\boldsymbol{\theta}), t \in [0,1]$ that is parameterized by $\boldsymbol{\theta}$.

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{t\sim U(0,1)} \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim \mathcal{D}_0} \sum_{i\in I} \max_{\mathrm{Dist}_i(\mathbf{x}',\mathbf{x})\leq \delta_i} \mathcal{L}(\phi(t;\boldsymbol{\theta});(\mathbf{x}',\mathbf{y})), \tag{4}$$

where the inner maximization part represents generating perturbed data from an adversarial strategy with distance measurement function Dist_i, and $I = \{1,2,\cdots,S\}$ denotes the number set of the considered types of adversarial strategies. For instance, $\mathbf{x}'$ can be produced by using ℓ_2 or ℓ_∞ norm distance measures, which are commonly used in adversarial attacks and adversarial training. Though out this paper, we restrict the size of I to be two, where the two adversarial strategies are all ℓ_p-PGD attacks and are the same as the AT used in two endpoints. We remark that the size of I can be larger and beyond the adversarial strategies used in two endpoints. For the two endpoints, we use two models trained by (3), possibly under different perturbation types, denoted as $\phi(0;\boldsymbol{\theta})$ and $\phi(1;\boldsymbol{\theta})$, respectively. In this paper, we adopt a quadratic Bezier curve to represent the path, whereby a model at a point t can be expressed as $\phi(t;\boldsymbol{\theta}) = (1-t)^2\boldsymbol{\theta}_1 + 2t(1-t)\boldsymbol{\theta} + t^2\boldsymbol{\theta}_2$. Similar to vanilla mode connectivity, (4) is a computationally feasible relaxation obtained by directly sampling t from a uniform distribution $U(0,1)$ during optimization. Data points $(\mathbf{x}',\mathbf{y})$ are generated by taking a union of adversarial strategies. Therefore, we ensure that the path we discover consistently adapts to the relevant adversarial perturbations.

We refer to (4) as the Robust Mode Connectivity (RMC). It is noteworthy that a group of models, comprising all points in the path, are produced from two initial models. Hence, RMC is a population-based optimization method. The subsequent step involves determining how to tackle (4).

3.2. Solving Robust Mode connectivity

Resolving (4) is a challenging task due to the presence of multi-type perturbations. The most straightforward approaches involve utilizing the 'MAX' or 'AVG' strategies proposed in [22], wherein the inner loss is acquired by selecting the type of perturbation that yields the maximum loss or by averaging the loss across all perturbation types. Nonetheless, both strategies treat perturbations independently. To overcome this limitation, we employ a Multi Steepest Descent (MSD) technique that includes diverse perturbation models within each step of the projected steepest descent to generate a PGD adversary with complete knowledge of the perturbation region [18]. The core concept involves maximizing the worst-case loss across all perturbation models at each step simultaneously.

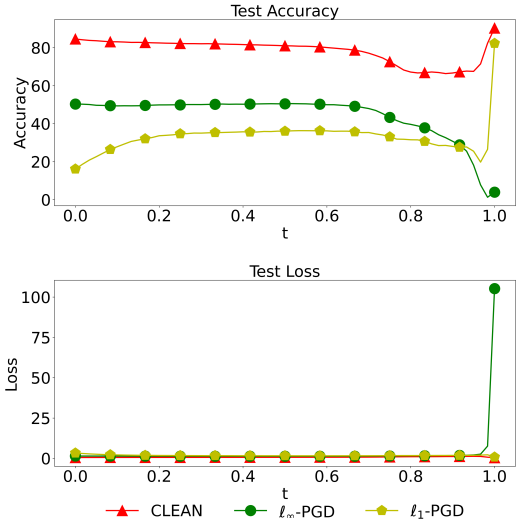
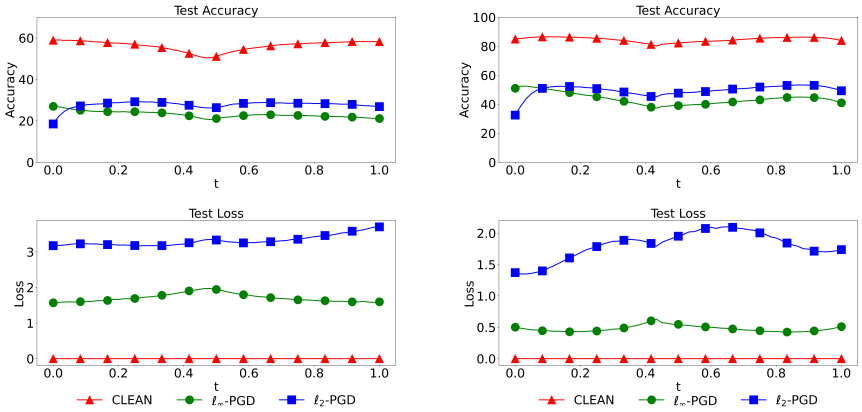


Figure 5. Curves obtained from two models trained by ℓ_1-AT and ℓ_∞-AT contain points with higher robustness on ℓ_1 and ℓ_∞-PGD attacks. The paths are obtained by training 50 epochs. The two endpoints are trained for 150 epochs.



(a) CIFAR100 (b) WideResNet-28-10

Figure 6. RMC is capable of discovering paths that contain points with high robustness on a range of datasets and model architectures. As shown in the left (right) figure, RMC finds a path with high robustness on the CIFAR-100 dataset (WideResNet-28-10 architecture). These paths were obtained by training for 50 epochs, with two endpoints trained using ℓ_∞ and ℓ_2-AT.

The details of the proposed RMC is shown in Algorithm 1. In each iteration, we consider all types of perturbations. To illustrate the effective of the proposed RMC, we conduct the following experiments. We utilize the two endpoints $\phi(0;\boldsymbol{\theta})$ and $\phi(1;\boldsymbol{\theta})$ trained with ℓ_∞-AT ($\delta = 8/255$, 150 epochs) and ℓ_2-AT ($\delta = 1$, 150 epochs) as before, and apply RMC (4) with MSD as the inner solver to obtain the path. Figure 4 displays the results of training for an additional 50/100/150 epochs using perturbed data generated by ℓ_2 and ℓ_∞ norm distance measures. In contrast to Figure 3, the paths contain points with both high accuracy and robustness against ℓ_∞-PGD and ℓ_2-PGD attacks. Although the left endpoint has low ℓ_2 robustness and the right endpoint has relatively low ℓ_∞ robustness, the points in the connection have larger ℓ_2 robustness (ℓ_∞ robustness) than the left (right) endpoint. In other words, robust mode connectivity can find a path with high robustness against all considered perturbations. Note that RMC is also a defense mechanism, as we can select the model in the path with the highest robustness. In the multi-type perturbation setting, the robustness is defined as the smallest robust accuracy under each ℓ_p-PGD attacks. The highest robustness is 48.19% in panel (a). Increasing the epoch number for solving (4) leads to smoother paths. Additionally, the optimal points in panels (a), (b), and (c) have similar robustness. If the goal is to select an optimal model from the path, conducting training with a small number of epochs is sufficient.

4. Experiments

The figures presented in Figures 3 and 4 demonstrate that utilizing the proposed Robust Mode Connectivity (RMC) can identify a path containing points that exhibit high robustness against various ℓ_p perturbations. In this section, we perform further experiments to provide a more thorough analysis of the effectiveness of RMC.

4.1. Settings

We conducted experiments to validate our proposed methods on CIFAR-10 and CIFAR-100 datasets [15], using PreResNet110 and WideResNet-28-10 architectures. In this work, we considered three types of perturbation norms, namely ℓ_∞, ℓ_2, and ℓ_1, with perturbations constrained by $\delta = 8/255$, 1, and 12, respectively, and we use AT to obtain endpoint models. We compared our methods with the standard ℓ_∞-AT baseline [17] and the state-of-the-art method MSD [18]. The evaluation metrics included standard accuracy on clean test data, robust accuracies under ℓ_∞, ℓ_2, ℓ_1-PGD adversarial attacks, and accuracy on worst-case sample-wise (Union) using all three basic PGD adversarial attacks, and robustness on ℓ_∞ and ℓ_2-PGD adversarial attacks.

4.2. Results

ℓ_∞-AT trained model with ℓ_1-AT trained model. We expanded our analysis by incorporating an additional ℓ_1-AT trained model, which we combined with the ℓ_∞-AT trained model. Figure 5 presents the results of our evaluation. We trained two endpoints for 150 epochs and obtained the path by conducting an additional 50 epochs. The right endpoint, corresponding to the ℓ_1-AT trained model, shows high resilience against ℓ_1 perturbations but is vulnerable to ℓ_∞ perturbations. Conversely, the left endpoint, i.e., the ℓ_∞-AT trained model, demonstrates high resilience against ℓ_∞ perturbations and can also withstand a certain level of ℓ_1 perturbations. By utilizing the RMC method, we obtained a

	Standard Accuracy	ℓ_∞-PGD ($\delta = 8/255$)	ℓ_2-PGD ($\delta = 1$)	ℓ_1-PGD ($\delta = 12$)	Union
ℓ_∞-AT	85.00%	49.03%	29.66%	16.61%	21.85%
MSD (**two** types of pert)	81.61%	48.57%	<u>45.92%</u>	35.64%	34.37%
RMC (ours, **two** types of pert)	80.90%	<u>48.19%</u>	48.63%	**38.05%**	**36.3%**

Table 1. RMC can achieve the highest robustness level against ℓ_∞-PGD and ℓ_2-PGD attacks among all defenses. We mark the robustness (the lowest robust accuracy) using an underline. RMC also has the highest robustness against the ℓ_1-PGD attack and Union. The baselines (ℓ_∞-AT [17] and MSD [18]) are trained with 200 epochs. Our RMC method with ℓ_∞ and ℓ_2 norm perturbations was trained with 150 epochs' endpoints and 50 epochs' path search.

path that exhibited improved robustness against both types of attacks.

RMC on CIFAR-100 and WideResNet-28-10. In this study, we assess the effectiveness of RMC using the WideResNet-28-10 model architecture on CIFAR-100. We explore two types of perturbations generated from ℓ_∞ and ℓ_2-PGD attacks. Our training approach includes 150 epochs for endpoints and an additional 50 epochs for path search costs. Figure 6 indicates that high robustness points are obtained when replacing CIFAR-10 with CIFAR-100 and PreResNet110 with WideResNet-28-10, demonstrating the adaptability of RMC across different datasets and architectures.

Comparing RMC with baseline methods. In Table 1, we provide a comparison between our proposed RMC, a ℓ_∞-AT baseline [17], and MSD [18]. According to the report from a recent work [4], [18] is the existing SOTA method for achieving diversified adversarial robustness. In this comparison, we are interested not only in identifying a robust path but also in pinpointing the optimal point. The optimal point is selected by checking the points with highest robustness on the curve. All baselines are trained for 200 epochs. We evaluated their performance under ℓ_∞-PGD and ℓ_2-PGD attacks, and marked the corresponding robustness (i.e., the lowest robust accuracy) with an underline. We also test these methods using the ℓ_1-PGD attack, which serves as a metric to evaluate the robustness against unforeseen attacks. Additionally, we highlighted the highest accuracy in the Union column. One can see that RMC can achieve the highest robustness level against ℓ_∞-PGD and ℓ_2-PGD attacks among all defenses. RMC also has the highest robustness against the ℓ_1-PGD attack and Union.

5. Conclusion

In this paper, we proposed a new method called robust mode connectivity (RMC) to search parameter space to find paths that enhance the adversarial robustness of neural networks (NNs) against multiple types of perturbations. By building on the concept of mode connectivity and employing a multi steepest descent (MSD) algorithm, RMC explored a wider range of parameter space and identified a path of NNs with high robustness against different types of ℓ_p norm perturbations ($p = 1, 2, \infty$). Extensive experiments on various datasets demonstrated the effectiveness of RMC in achieving diversified adversarial robustness and outperforming existing adversarial training methods.

Our work suggests that exploring a broader range of parameter space is crucial for achieving strong adversarial robustness in NNs. The proposed RMC method provides a new avenue for enhancing adversarial robustness by leveraging population-based optimization. Future work can investigate the extension of RMC to more types of perturbations and evaluate its robustness against more complex attacks. To further improve the robustness, it is possible to develop a multi-stage RMC-based optimization regime that contains multiple RMC units and takes different perturbation types into consideration. The proposed method can also be applied to other machine learning tasks beyond adversarial robustness, where exploring a wider range of parameter space can enhance model performance.

References

[1] Jianing Bai, Ren Wang, and Zuyi Li. Physics-constrained backdoor attacks on power system fault localization. *IEEE PES General Meeting*, 2023. 1

[2] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017. 1

[3] Francesco Croce and Matthias Hein. Provable robustness against all adversarial ℓ_p-perturbations for $p \geq 1$. In *International Conference on Learning Representations*, 2019. 1

[4] 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*, pages 4436–4454. PMLR, 2022. 1, 6

[5] Xiaodong Cui, Wei Zhang, Zoltán Tüske, and Michael Picheny. Evolutionary stochastic gradient descent for optimization of deep neural networks. *Advances in neural information processing systems*, 31, 2018. 2

[6] Zhongying Deng, Xiaojiang Peng, Zhifeng Li, and Yu Qiao. Mutual component convolutional neural networks for hetero-

geneous face recognition. *IEEE Transactions on Image Processing*, 28(6):3102–3114, 2019. 1

[7] Rida T Farouki. The bernstein polynomial basis: A centennial retrospective. *Computer Aided Geometric Design*, 29(6):379–419, 2012. 2

[8] C Daniel Freeman and Joan Bruna. Topology and geometry of half-rectified network optimization. 2016. 2

[9] 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*, 31, 2018. 2

[10] Jonas Gomes, Luiz Velho, and Mario Costa Sousa. *Computer graphics: theory and practice*. CRC Press, 2012. 2

[11] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016. 3

[13] Yuge Huang, Jiaxiang Wu, Xingkun Xu, and Shouhong Ding. Evaluation-oriented knowledge distillation for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18740–18749, 2022. 1

[14] Max Jaderberg, Valentin Dalibard, Simon Osindero, Wojciech M Czarnecki, Jeff Donahue, Ali Razavi, Oriol Vinyals, Tim Green, Iain Dunning, Karen Simonyan, et al. Population based training of neural networks. *arXiv preprint arXiv:1711.09846*, 2017. 2

[15] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009. 5

[16] 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. 1

[17] 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. 1, 3, 5, 6

[18] Pratyush Maini, Eric Wong, and Zico Kolter. Adversarial robustness against the union of multiple perturbation models. In *International Conference on Machine Learning*, pages 6640–6650. PMLR, 2020. 1, 2, 4, 5, 6

[19] Saroj Kumar Pandey and Rekh Ram Janghel. Recent deep learning techniques, challenges and its applications for medical healthcare system: a review. *Neural Processing Letters*, 50:1907–1935, 2019. 1

[20] 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*, pages 3353–3364, 2019. 1

[21] David Stutz, Matthias Hein, and Bernt Schiele. Confidence-calibrated adversarial training: Generalizing to unseen attacks. In *International Conference on Machine Learning*, pages 9155–9166. PMLR, 2020. 1

[22] Florian Tramer and Dan Boneh. Adversarial training and robustness for multiple perturbations. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 4

[23] Jingkang Wang, Tianyun Zhang, Sijia Liu, Pin-Yu Chen, Jiacen Xu, Makan Fardad, and Bo Li. Adversarial attack generation empowered by min-max optimization. *Advances in Neural Information Processing Systems*, 34:16020–16033, 2021. 1

[24] Ren Wang, Tianqi Chen, Stephen Lindsly, Cooper Stansbury, Indika Rajapakse, and Alfred Hero. Immuno-mimetic deep neural networks (immuno-net). *The 2021 ICML Workshop on Computational Biology*, 2021. 2

[25] 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*, 10:22061–22078, 2022. 2

[26] Ren Wang, Tianqi Chen, Philip Yao, Sijia Liu, Indika Rajapakse, and Alfred O Hero. Ask: Adversarial soft k-nearest neighbor attack and defense. *IEEE Access*, 10:103074–103088, 2022. 1

[27] 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. 1