Implicit Euler Skip Connections: Enhancing Adversarial Robustness via Numerical Stability

Anonymous Authors[1]

Abstract

Deep neural networks have achieved great success in various areas. However, recent works have found that neural networks are vulnerable to adversarial attacks, which leads to a hot topic nowadays. Although many approaches have been proposed to enhance the robustness of neural networks, few of them explored robust architectures for neural networks. On this account, we try to address such an issue from the perspective of dynamic system in this work. By viewing ResNet as an explicit Euler discretization of an ordinary differential equation (ODE), for the first time, we find that the adversarial robustness of ResNet is connected to the numerical stability of the corresponding dynamic system. Namely, more stable numerical schemes may correspond to more robust deep networks. Furthermore, inspired by the implicit Euler method for solving numerical ODE problems, we propose Implicit Euler skip connections (IE-Skips) by modifying the original skip connection in ResNet or its variants. Then we theoretically prove its advantages under the adversarial attack. Experimental results show that our ResNet with IE-Skips can largely improve the robustness and the generalization ability under adversarial attacks when compared with the vanilla ResNet of the same parameter size.

1. Introduction

Deep Learning (DL) has achieved great success in many machine learning problems and has been widely used in various computer vision and neural language processing tasks. However, recent works show that Neural Networks are vulnerable to adversarial attacks (Zügner et al., 2018; Moosavi-Dezfooli et al., 2016; Szegedy et al., 2013), i.e., adding some human-imperceptible noise to the input may lead to incorrect predictions even for the state-of-the-art convolution neural networks. Such defect significantly hampers the applications to the safety-critical problems such as those in medical diagnosis.

In order to handle the problems above, researchers have proposed many training methods to make the network more robust recently, such as the Projected Gradient Descent (Madry et al., 2017) method, YOPO (Zhang et al., 2019a), TRADES (Zhang et al., 2019b), etc.. But few of them discussed the immanent robustness of a network structure. In this paper, we concentrate on finding a more robust and powerful neural architecture via our theoretical analysis.

There have been many works that bridge the dynamic systems and the neural networks especially ResNet (He et al., 2016a). Under the assumptions that ResNet is a kind of explicit (forward) Euler discretization, lots of ResNet variants (Zhu et al., 2018) are proposed based on more accurate low-order numerical schemes for solving the corresponding ODEs. While the existing works are based on the assumption that higher natural accuracy (i.e., the accuracy without adversarial attacks) corresponds to more *accurate* numerical schemes, we are the first to propose that higher robust accuracy (i.e., the accuracy under adversarial attacks) corresponds to more *stable* numerical schemes. This is because they are both the problem of how sensitive the output is to the small perturbation of the input.

As shown in Figure 1, the stability of the explicit Euler method is weak, i.e. a small change on the initial point leads to a tremendous variation of the output. Accordingly, the robustness of ResNet, which corresponds to the explicit Euler scheme, is not satisfactory under adversarial attacks. However, it is widely known that the stability of the implicit (backward) Euler discretization is outstanding, as shown in Figure 1. As the stability of the implicit method is superior to the explicit ones in numerical ODE, we propose an implicit-Euler architecture by unfolding the implicit Euler method. The architecture can be utilized in any networks with skip connections. If we use it in the vanilla ResNet, we can obtain a stable network called IE-ResNet which gains large improvements on the robustness of the network with no more parameter consumption.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

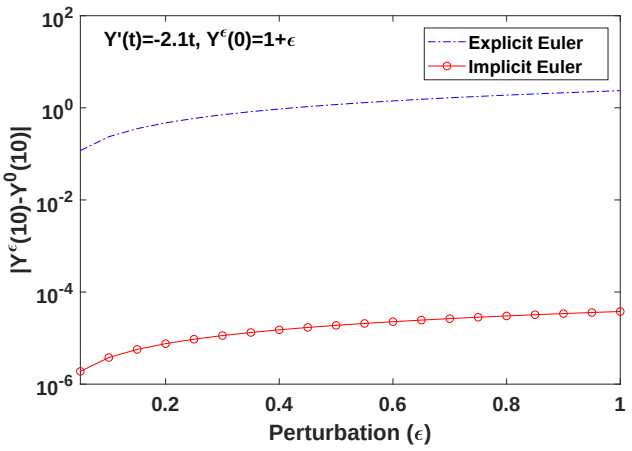

Figure 1. The stability of an initial value problem with the explicit (forward) and the implicit (backward) Euler schemes.

Furthermore, we theoretically prove that our IE-ResNet can be resistant to adversarial attacks with higher probability than the vanilla ResNet. We also conduct various experiments to verify the robustness of our IE-ResNet and our analysis.

2. Related Works and Contributions

2.1. Related Works

Adversarial Defense. Many works have been proposed to enhance the robustness of neural networks, such as adversarial training and its variants (Madry et al., 2017; Zhang et al., 2019a; Shafahi et al., 2019), various regularizations (Cisse et al., 2017; Lin et al., 2019; Jakubovitz & Giryes, 2018), generative model based defense (Sun et al., 2019), Bayesian adversarial learning (Ye & Zhu, 2018), TRADES method (Zhang et al., 2019b), etc. All the methods above aim to improve the robustness of the networks via training by different strategies. Apart from that, some works propose some robust architectures like adding noise via an intriguing stochastic differential equation perspective (Wang et al., 2018) or feature denoising (Svoboda et al., 2018; Xie et al., 2019). However, none of them studies the inherent robustness of the ResNet architecture from the stability view of dynamic systems and proposes a more robust one via dynamic system perspectives. Therefore, in this paper, we aim to propose a new robust architecture which can improve the robustness of the original Residual Network family.

Neural Networks and ODEs. Many works these days have bridged the relationship between the ODEs and neural networks (Chen et al., 2018; Lu et al., 2017). Especially for the ResNet, researchers have found that it can be written as the explicit Euler discretization with unit steps. From the above view, many training methods (Li & Hao, 2018) have been proposed to train neural networks via the optimal control perspective. In addition to that, many new architectures (Haber & Ruthotto, 2017; Zhang et al., 2019c; Yang et al., 2019) have been proposed to improve their performance, inspired by more accurate numerical ODE methods. However, none of them explores the adversarial robustness under the perspective of the stability of the numerical ODE or dynamic system.

2.2. Contribution

- To the best of our knowledge, this is the first work to consider the adversarial robustness of a neural network from the perspective of the dynamic system stability. From such an aspect, we explore the impact of dynamic stability of ResNet on its adversarial robustness.
- Building on the above insight, we propose the IE-Skips, which modifies the original skip connections in the Residual Network family inspired by the implicit Euler discretization. We theoretically prove that ResNets with our IE-Skips (called IE-ResNet) is more robust against the adversarial attack than vanilla ResNet.
- On the MNIST and the CIFAR benchmarks, we conduct experiments to verify the adversarial robustness of IE-ResNets, which replace the original skip connections with our IE-Skips in ResNets. The experimental advantages demonstrate the robustness of our architecture and validate our analysis based on dynamic systems.

3. Network Robustness and Stability of Dynamic Systems

3.1. Preliminaries and Notations

We use $(\mathbf{x}_0, \mathbf{y}_0)$ to denote a pair of input and label for training or testing. $\mathcal{F}(\mathbf{x})$ represents the output of the network. For a function $f : \mathbb{R}^d \to \mathbb{R}^d$, we use $\nabla f(\mathbf{x})$ to denote its Jacobian at input $\mathbf{x}$. We let $\mathbb{B}^{(n)}(\mathbf{x}, r)$ denote n-dimensional ball centered at $\mathbf{x}$ with radius r. We call an N-stage network if the output $\mathbf{x}_N$ and input $\mathbf{x}_0$ of the network abide by the following equation:

$$\mathbf{x}_i = s_i(\mathbf{x}_{i-1}), \text{ for } i = 1, ..., N,$$

where we call $s_i(\cdot)$ the i-th stage of the network.

3.2. Stability of Dynamic Systems and Network Robustness

Recent works have built the relationship between the dynamic systems and neural networks and proposed many architectures and training methods with better generaliza-

tion ability, inspired by some numerical algorithms for dynamic systems. Furthermore, some researchers interpret new ResNet variants from such a perspective, where the i-th residual stage is formulated as follows:

$$\mathbf{x}_i = \mathbf{x}_{i-1} + hf_i(\mathbf{x}_{i-1}),\ i = 1, ..., N,$$

in which $h \leq 1$ is a constant (Zhang et al., 2019c) or a learnable parameter (Yang et al., 2019). (Zhang et al., 2019c; Yang et al., 2019) both demonstrate that a smaller h can better control the output variation under the Guassian noise injection, which is consistent with the theoretical results of the choice of h on the stability of the explicit Euler method. Such an observation motivates us to explore the links between the stability of dynamic systems and the robustness of its corresponding networks under the adversarial attacks.

As we will illustrate in the following, we find that the numerical stability on the initial value problem is similar to the network's robustness against adversarial attacks which add perturbations to the input, especially when training the network with least squared regression loss. First of all, we define the numerical stability for an N-stage neural network from the dynamic system perspective as follows:

Definition 1. *A network with N stages (s_i represents its i-th stage) is called C-stable for its initial value problem at input $\mathbf{x}_0 \in \mathbb{R}^n$, if for a small δ and all the perturbed inputs for each stage $\mathbf{x}'_{i-1} \in \mathbb{B}^{(n)}(\mathbf{x}_{i-1}, \delta)$, the following equations are satisfied for all the stages:*

$$\|s_i(\mathbf{x}'_{i-1}) - s_i(\mathbf{x}_{i-1})\|_2 \leq C\delta,\ \text{for } i = 1, ..., N,$$

where $C \leq 1$ is a constant.

From the above definition, one can see that if the network is C-stable at certain input $\mathbf{x}_0$, then the impacts of the small adversarial perturbation will not enlarge, or even shrink, during the forward propagation. Furthermore, we can bound the increment of loss for any attacks $\boldsymbol{\eta} \in \mathbb{B}^{(n)}(0, \delta)$ for sample $\mathbf{x}_0$ if the network is C-stable at $\mathbf{x}_0$ using the least squared regression loss.

Proposition 1. *If a network with N stages is C-stable at $\mathbf{x}_0$, then the increment of the least squared regression loss under the adversarial attack $\boldsymbol{\eta} \in \mathbb{B}^n(0, \delta)$ on input $\mathbf{x}_0$ is:*

$$\mathcal{L}(\mathcal{F}(\boldsymbol{\theta}; \mathbf{x}_0 + \boldsymbol{\eta}), \mathbf{y}_0) - \mathcal{L}(\mathcal{F}(\boldsymbol{\theta}; \mathbf{x}_0), \mathbf{y}_0) \leq C^N\delta,$$

where $\mathcal{F}(\cdot)$ denotes the neural network and $\mathbf{y}_0$ is the label for clean data $\mathbf{x}_0$.

Therefore, if the network is C-stable at sample $\mathbf{x}_0$, then the impacts of a small perturbation $\boldsymbol{\eta}$ on the output will remain the same or become smaller during the inference no matter suffering what kind of attack. Consequently, the network can perform more stably under the attacks on such samples. We call that a network can defend the adversarial attacks on $\mathbf{x}_0$ if the network is C-stable with $C \leq 1$ in the following analysis.

3.3. ResNet and Its Robustness Conditions

Neglecting the size and the dimension change on the input, the forward propagation of the input value $\mathbf{x}_0$ for a vanilla ResNet (He et al., 2016a) with N residual stages can be depicted as follows:

$$\mathbf{x}_i = \mathbf{x}_{i-1} + f_i(\mathbf{x}_{i-1}),\ i = 1, ..., N, \tag{1}$$

where $f_i(\cdot)$ denotes the i-th residual block with $\mathbf{x}_{i-1}$ and $\mathbf{x}_i$ representing its input and output, respectively. Therefore, ResNet can be considered as an explicit Euler discretization with unit step size for the following initial value problem (E, 2017; Haber & Ruthotto, 2017):

$$\dot{\mathbf{x}}(t) = f_t(\mathbf{x}(t)),\ \mathbf{x}(0) = \mathbf{x}_0,$$

where features $\mathbf{x}(t)$ are viewed in the continuous limit as a function of time $t \in [0, N]$. With the above insights, we analyze the sufficient conditions that ResNet and our network can defend the adversarial attacks.

Proposition 2. *For an N-block Residual Neural Network with f_i representing its i-th residual block and a small $\delta > 0$, if the following statement is satisfied:*

$$\|\mathbf{I} + \nabla f_i(\mathbf{x}_{i-1})^\top\|_2 \leq 1 \text{ for } i = 1, ..., N, \tag{2}$$

where $\mathbf{x}_{i-1}$ denotes the input of the i-th block corresponding to the clean input $\mathbf{x}_0$ for the network, then the network with N blocks can defend the attack with perturbation $\boldsymbol{\eta} \in \mathbb{B}^{(n)}(0, \delta)$ on the sample $\mathbf{x}_0$.

As one can see from the above proposition, such conditions for the Jacobian of ResNets are not easy to satisfy for an input $\mathbf{x}_0$ during training, let alone for practical data points sampled from some distribution. On this account, the vanilla ResNet is sensitive to adversarial perturbations.

4. IE-Skips and Its Robustness Analysis

4.1. IE-Skips Architecture

Although the explicit Euler method is a popular first-order approach for solving ODEs numerically, its stability conditions (illustrated in Prop. 2) for an input $\mathbf{x}_0$ are hard to realize. For this reason, the vanilla ResNet often fails at different attacks. From the existing theory on numerical methods for ODEs, implicit Euler is a well-known first-order method with sound stability. On this account, we aim to revise the original ResNet to enhance its robustness with the implicit Euler method.

Implicit Euler is a widely used method in numerical ODE with the following equation:

$$\mathbf{y}_i = \mathbf{y}_{i-1} + hg(\mathbf{y}_i), \tag{3}$$

where h stands for the step size. Current implicit Euler method often uses fixed-point method or gradient methods to solve Eqn (3). Since we cannot ensure that the residual block always performs as a contractive mapping, we choose to use the gradient descent method to approximate the implicit scheme in our network structure. Like ResNet, we set $h = 1$ in our network and integrate a non-linear least square optimization procedure via gradient descent into the original Skip-Connection to form a new architecture called Implicit Euler Skips connections (IE-Skips). The details for the residual stage with IE-Skips are illustrated in Alg. 1.

Algorithm 1 Forward Propagation of the i-Residual Stage with Our IE-Skips.

Input: Input from the former stage $\mathbf{x}_{i-1}$, the residual block $f_i(\cdot)$, inner iteration number K and inner step size γ (We choose to be 0.05 or 0.1 in our experiments).

1: Compute $\mathbf{x}_i = \mathbf{x}_{i-1} + f_i(\mathbf{x}_{i-1})$.
2: **for** $k = 1$ to K **do**
3: Compute $r_i = \|\mathbf{x}_i - \mathbf{x}_{i-1} - f_i(\mathbf{x}_i)\|_2^2$.
4: Compute $\nabla_{x_i} r_i$.
5: Update $\mathbf{x}_i$ via

$$\mathbf{x}_i = \mathbf{x}_i - \gamma \nabla_{x_i} r_i.$$

6: **end for**

Output: The output of the i-residual stage $\mathbf{x}_i$.

Note that in Alg. 1, we use the original outputs of the vanilla residual stage (listed in step 1 of Alg. 1) as the initial point for the nonlinear least square problem. In this way, IE-Skips can preserve the merits of the original skip connection and ensure the original advantages for gradient back-propagation. Then with a few steps of gradient descent, we can rectify the vanilla residual stage to our Implicit Euler residual stage with almost no harm to its representation ability. With our IE-Skips, we can modify all networks in the Residual Network family to improve their robustness by replacing the original residual skip connections with our IE-Skips when the input and the output are of the same dimension and size (as shown in Fig. 2). In this way, we construct a new model called IE-ResNet which is more robust than the vanilla ResNet.

Like ResNet, we theoretically analyze the superiority of the stability of our IE-ResNet over the original ResNet in the following.

4.2. Stability of Our IE-ResNet

First of all, we theoretically analyse the robustness condition for our exact IE-ResNet, which solves the non-linear square least problem exactly, like Prop. 2 for ResNet.

Proposition 3. *For an N-block exact IE-ResNet with f_i representing its i-th residual block and a small $\delta > 0$, if the following statement is satisfied:*

$$\sigma_{min}(\mathbf{I} - \nabla f_i(\mathbf{x}_i)^\top) \geq 1 \text{ for } i = 1, ..., N, \quad (4)$$

where σ_{min} denotes the smallest singular value and $\mathbf{x}_i$ denotes the output of the i-th block corresponding to the clean input $\mathbf{x}_0$ for the network, then the network with N blocks can defend attacks with perturbation $\boldsymbol{\eta} \in \mathbb{B}^{(n)}(0, \delta)$ on sample $\mathbf{x}_0$.

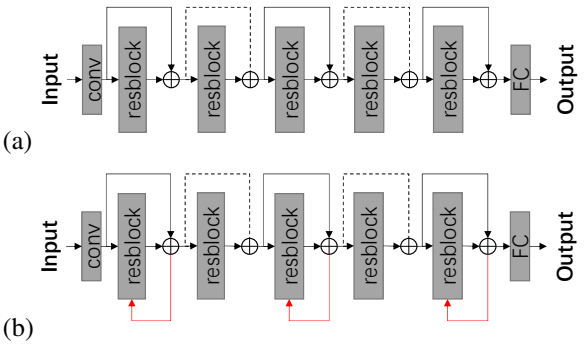

Figure 2. The structure sketch of the vanilla ResNet (a) and our ResNet with IE-Skips (b), which is called IE-ResNet. The solid black lines in (a) denote the skip connections while the dotted black lines represent the dimension changing operator (we use 3×3 convolutions in the following). The soild black lines combined with red lines in (b) represent the IE-Skips and the red lines in (b) denote the non-linear least square optimization process (Line 3-5 in Alg. 1).

With the propositions above, we are going to prove that our IE-ResNet has higher probability to defend the attack under our definitions above than its corresponding ResNet in the following theorem.

For an N-block ResNet, we use g_i to represent its i-th residual block. Meanwhile, for an N-block exact IE-ResNet, we use f_i to represent its i-th residual block. Furthermore, we use $\mathbf{x}_i$ to denote the input of the i-th block of ResNet while we use $\mathbf{y}_i$ to represent the output of the i-th block of IE-ResNet.

Theorem 1. *Suppose that for an input $\mathbf{x}$, which is sampled from a data distribution, its corresponding $\nabla g_i(\mathbf{x}_i)$ and $\nabla f_i(\mathbf{y}_i)$ obey the same distribution since they enjoy the same strategies and Jacobians $\{\nabla g_i(\mathbf{x}_i), \nabla f_i(\mathbf{y}_i)\}$ are independent. Then, we can obtain the following relations:*

$$\mathbb{P}[\cap_{i=1,..N}\{\|\mathbf{I} + \nabla g_i(\mathbf{x}_i)^\top\|_2 \leq 1\}] \leq \mathbb{P}[\cap_{i=1,..N}\{\sigma_{min}(\mathbf{I} - \nabla f_i(\mathbf{y}_i)^\top) \geq 1\}].$$

From the above theorem, one can see that the possibility for our IE-ResNet to maintain stability on a sample is higher than that for the vanilla ResNet. So the robustness of our IE-ResNet is superior to the vanilla ResNet under above

definitions and assumptions. Although our analysis above only concentrate on the ideal circumstances, the experimental results in the following also demonstrate the merits of our IE-Skips and confirm that our conclusions can be also extended to practical problems.

In addition, our implicit scheme can be integrated into many ResNet-like networks such as ResNeXt. For the sake of convenience, we use IE-ResNet-K-d to represent our d-layer ResNet with K-inner-iteration IE-Skips. Although the complexity for training our model is about $K+1$ times than before, the model size is the same as the vanilla ResNet. In the following sections, we conduct experiments for our IE-ResNet-K-d and ResNet-d on different datasets to validate our advantages over the vanilla ResNets.

5. Experiments

5.1. Adversarial Attack

In this subsection, we first introduce three popular white-box and black-box attack methods: Projected Gradient Descent (PGD) (Madry et al., 2017), Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2014) and Carlini & Wagner (C&W) (Carlini & Wagner, 2017). In the following expeirments, we use the above three attack methods to evaluate the robustness of different models and demonstrate the advantages of our model.

C&W Attack C&W attack searches the targeted adversarial image by solving the following optimization problem:

$$\begin{aligned} &\min_{\boldsymbol{\eta}} \|\boldsymbol{\eta}\|_2, \\ &s.t. \mathcal{F}(\mathbf{w}, \mathbf{x}+\boldsymbol{\eta}) = t, \\ &\mathbf{x}+\boldsymbol{\eta} \in [0,1]^d, \end{aligned}$$

where δ is the adversarial perturbation and t is the target label.

PGD Attack PGD searches the adversarial instances $\mathbf{x}'$ by iteratively increasing the loss function $\mathcal{L}(\mathbf{x}', y) = \mathcal{L}(f(\mathbf{x}', \mathbf{w}), y)$, subject to the constraint $\|\mathbf{x}' - \mathbf{x}\|_\infty \leq \delta$ where $\mathbf{x}$ is the clean instance and $\|\cdot\|_\infty$ is the infinite norm. Using the projected gradient descent method with step size α, the perturbed instances can be generated:

$$\mathbf{x}^{(m)} = \text{Clip}_{\mathbf{x},\delta}\{\mathbf{x}^{(m-1)} + \alpha \cdot \text{sign}(\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^{(m-1)}, y))\},$$

where $m = 1, ..., M$, $\mathbf{x}^{(0)} = \mathbf{x}$ and $\text{Clip}_{\mathbf{x},\delta}\{\cdot\}$ clips the input to make the output lies in the ℓ_∞ ball around $\mathbf{x}$ with radius δ. $\mathbf{x}' = \mathbf{x}^{(M)}$ is the adversarial image generated by PGD-M attack. Furthermore, if we set $\alpha = \delta$ and $M = 1$, then we can obtain the Fast Gradient Sign Method (FGSM), which is also a widely used adversarial attack approach.

In the following experiments, we set $\delta = 8/255$ and $\alpha = 1/255$ on CIFAR if we use the PGD method for adversarial training and evaluation. In addtion to these methods, we also run Adam for 50 iterations with learning rate equaling 6×10^{-4} and $c = 10$ for C&W adversarial evaluation. As for MNIST, we set $\delta = 0.15$ for FGSM adversarial training or evaluation. We implement all the experiments with PyTorch (Paszke et al., 2019).

5.2. IE-ResNet with Natural Training

In this part, we conduct experiments on MNIST (LeCun & Cortes, 2010) and CIFAR-10 (Krizhevsky et al., 2009) to demonstrate the robustness of IE-ResNets. For natural training, we train the PreAct-ResNets (He et al., 2016b) and IE-ResNets with inner step size $\gamma = 0.05$ on clean data. Then we compare their performance on the datasets perturbed by the FGSM attack with $\delta = 0.15$ for MNIST and $\delta = 1/255$ for CIFAR-10, respectively.

MNIST First of all, we resize the vanilla ResNet-18 for ImageNet with 8 initial channels for MNIST. Then, we naturally train (i.e., without adding adversarial samples) the ResNet-18 and our model with the same size on MNIST for 35 epochs. Finally, we use white-box FGSM attack to evaluate their robustness. The results are shown in Table 1. Natural Accuracy here means the accuracy for the model evaluated on the clean datasets and Robust Accuracy means the accuracy for the model evaluated on the perturbed datasets by adversarial attacks.

Models	Natural Acc(%)	Robust Acc(%)
ResNet-18	99.41	88.45
IE-ResNet-1-18	99.32	91.36
IE-ResNet-3-18	99.38	92.27

Table 1. Natural Accuracy and Robust Accuracy for IE-ResNet-18 and ResNet-18 on MNIST by natural training. The adversarial attack is FGSM attack.

From the results, one can see that our models achieve higher robust accuracy on the MNIST dataset with comparable predictive performance on the clean datasets. The results confirm our anlaysis that our IE-ResNet can be resistant to adversarial attacks with higher probability than the vanilla ResNet.

CIFAR-10 Besides small datasets, we also evaluate the robustness and generalization abilities for the naturally trained models (the initial channel number equals to 16) on CIFAR-10. Firstly, we naturally train IE-ResNets and ResNets with different depths for 180 epochs with initial learning rate 0.05, which decays by a factor 5 at the 80th, 120th and 160th epochs. Then we use the FGSM attack to evaluate the robustness of our model and the vanilla ResNet. The

results are shown in Table 2. Furthermore, Figure 3 plots the evolution of training and validation accuracies of different models.

Models	Natural Acc(%)	Robust Acc(%)
ResNet-50	91.65	67.89
IE-ResNet-1-50	92.22	68.69

Table 2. The Natrural Accuracy and Robust Accuracy for IE-ResNet-50 and ResNet-50 trained by clean data. The adversarial attack is the FGSM attack.

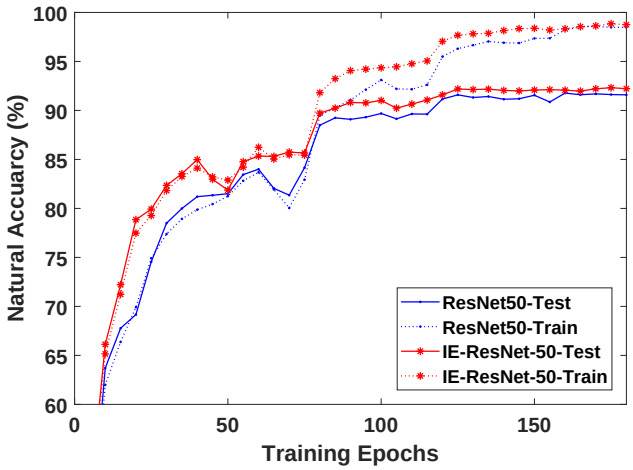

Figure 3. Evolution of training and validation accuracies for ResNet-50 and IE-ResNet-50.

From the tables and the figure, one can see that IE-ResNet can beat the vanilla ResNet not only under the natural evaluation but also under the robust evaluation. The empirical experiments confirm our analysis that IE-ResNet enjoys more stability and generalization ability than ResNet.

5.3. IE-ResNet with Robust Training

Since our model does not involve any constraints to the network's Jacobian, the exact robustness conditions in our theory are hard to satisfy with natural training. On this account, our model cannot successfully preserve stability if the attacks are strong via natural training although our IE-ResNet is more robust than the vanilla ResNet from both theoretical or experimental aspects. However, like the vanilla ResNet, we can utilize the adversarial training to enhance the robustness of our model. As we can see from the following results on CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009), our models can obtain higher robust accuracy than the vanilla ResNet with the same adversarial training method, which also demonstrates the robustness of our models. In this part, we use the models resized from ResNets for ImageNet, whose channel numbers are 4 times smaller. In the following experiments, we use PGD-10 adversarial training for both CIFAR-10 and CIFAR-100 with the same learning rate schedule and epochs as the natural training.

CIFAR-10 Firstly, we evaluate the robustness via the PGD-20 attack on the models with different depths shown in Table 3. From the results, one can see that the robust accuracies of our IE-ResNets are consistently higher than those of their corresponding vanilla ResNets, while having comparable predictive capabilities as those of the vanilla ResNets. Therefore, we can conclude that our modification can promote the robustness of the vanilla ResNets and validate our analysis. Besides, the improvements on IE-ResNet-1-18, IE-ResNet-1-34 (without bottleneck) and IE-ResNet-1-50 (with bottleneck) also show that our modification can be applied in various ResNet variants and enhance their robustness since no constraints on the residual blocks have been imposed in our analysis.

Models	Natural Acc (%)	Robust Acc(%)
ResNet-18	81.44	38.08
IE-ResNet-1-18	81.89	38.71
ResNet-34	82.59	40.11
IE-ResNet-1-34	83.57	41.76
ResNet-50	84.34	41.43
IE-ResNet-1-50	85.28	42.15

Table 3. Natural Accuracies and Robust Accuracies of IE-ResNets and ResNets on CIFAR-10. The attack for the robust evaluation is PGD-20 attack.

In addition to the popular PGD attack, we also evaluate the models with the C&W attack shown in Fig. 4. As one can see from the figure, our IE-ResNet is consistently more robust than the vanilla ResNet under the C&W attack.

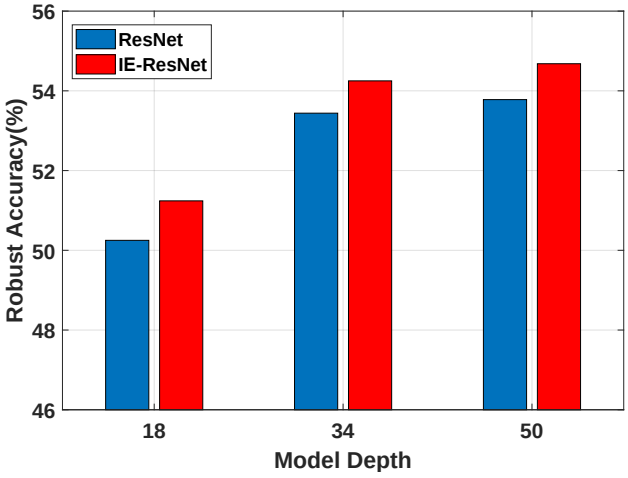

Figure 4. The robust accuracy of ResNet and our IE-ResNet (inner step size equals to 1) trained by PGD-10 with different depths under the C&W attack.

Furthermore, we also do experiments to evaluate the robustness of IE-ResNet and its corresponding ResNet under the black-box attack, as shown in Table 4. In this scenario, we use the target model to classify the adversarial images crafted by applying the FGSM attack ($\epsilon = 16/255$) to the oracle model as listed in the table. As we can see, our IE-ResNet also performs better than ResNet under the black-box adversarial attack, which also demonstrates the robustness of our model.

Models	Oracle	Robust Acc(%)
ResNet-50	IE-ResNet-1-50	38.26
IE-ResNet-1-50	ResNet-50	39.41

Table 4. Robust Accuracies of IE-ResNet and ResNet on adversarial images of CIFAR-10 crafted by attacking the oracle model with the FGSM attack.

CIFAR-100 Besides evaluating the robustness on CIFAR-10, we also compare the robustness of ResNet-50 and IE-ResNet-1-50 on the CIFAR-100 benchmark. First of all, we train both models via PGD-10 adversarial training and then evaluate their robustness performance under PGD attacks with different steps:$\{10, 20, 30, 50, 80, 100\}$. The curves of the robust accuracy for different models with respect to the PGD attack with various iterations are shown in Figure 5.

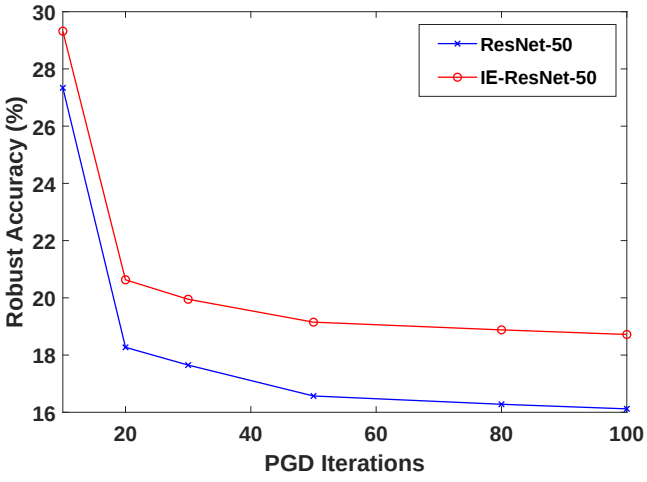

Figure 5. The robust accuracy of ResNet-50 and IE-ResNet-1-50 under the PGD attack with different number of iterations on CIFAR-100.

As one can see, our IE-ResNet-1-50 also performs much more robust under adversarial attacks on CIFAR-100 benchmark. Moreover, the robust accuracy gap between our IE-ResNet-50 and ResNet-50 gets larger with the attack going stronger. From the experiments above, we can conclude that IE-ResNets are much more robust than its corresponding ResNets under various circumstances.

5.4. IE-ResNet with TRADES

It is widely known that many improvements may somehow improve the performance of small networks but not be effective on large ones. However, according to our analysis, our architecture can consistently ameliorate the robustness of the network no matter the size of the model. In order to confirm that, in this section, we conduct the experiments on the widely-used WideResNet-34-10, which has 160 initial channels. Utilizing our architecture, we obtain our IE-WideResNet-1-34 (with inner step size $\gamma = 0.1$).

Apart from that, we use TRADES method (Zhang et al., 2019b) with $\lambda = 1/6$ for adversarial training. TRADES formulates a trade-off between robustness and accuracy as follows,

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}} \max_{\boldsymbol{\eta}\leq\delta} (\mathcal{L}(f_{\boldsymbol{\theta}}(\mathbf{x}, y) + \ell(f_{\boldsymbol{\theta}}(\mathbf{x}), f_{\boldsymbol{\theta}}(\mathbf{x}+\boldsymbol{\eta})/\lambda)),$$

where $f_\theta(\mathbf{x})$ is the neural network parameterized by $\boldsymbol{\theta}$, $\mathcal{L}$ represents the loss function, ℓ denotes the consistency loss and λ is a balancing hyper-parameter. If the neural network is trained by solving the above min-max problem, we call the network is trained by TRADES-$1/\lambda$ method.

Comparing the results of our reproduced WideResNet-34 and IE-ResNet-1-34 with the same setting as shown in Table 5, one can see that our model outperforms the vanilla WideResNet no matter sustaining what kind of attack with obvious advantages (around 3% improvement under different attacks).

Furthermore, as one can see from Table 5, our IE-Skips improve both the natural accuracy (85.62% vs. 84.92%) and robust accuracy (58.06% vs. 56.61%) for WideResNet-34 under PGD-20 attack comparing with the results reported by (Zhang et al., 2019b),which is the state-of-the-art results, with only 5% size larger than vanilla WideResNet (46M vs. 44M). The parameter increments are caused by the down-sampling connection we use. When comparing with another stable architecture En$_1$WideResNet-34 modified on WideResNet, our model enjoys a distinct advantage under the adversarial evaluation with comparable predictive capability on clean dataset, which demonstrates that the robustness of our model is better than En$_1$ResNet. From the experimental results, we can conclude that our network can not only perform well on the small networks, but also consistently improve the robustness of large networks even for the state-of-the-art ones.

6. Discussions

6.1. Impacts of the Inner Iterations

As we analyzed in Section 4.2, ResNets with exact implicit scheme is much more robust than the vanilla ResNet. Therefore, being closer to the exact implicit ResNet can make the

Models	Natural (%)	PGD-20 (%)	PGD-100 (%)	C&W (%)
WideResNet-34 (Madry et al., 2017)	87.30	47.04	–	–
En_1WideResNet-34 (Wang et al., 2018)	86.19	56.60	–	–
WideResNet-34 (Zhang et al., 2019b)	84.92	56.61	–	–
WideResNet-34 (Our implementation)	84.95	55.50	53.95	61.62
IE-WideResNet-1-34	**85.62**	**58.06**	**57.01**	**66.25**

Table 5. Comparisons of different models with different training methods under different adversarial attacks. The upper three results are copied directly from the papers and the hypen here means that the robust accuracies under such attacks are not stated in their papers.

network more robust. As more inner iterations in IE-Skips can make our IE-ResNet perform more similarly to the exact implicit scheme, we want to validate whether more iterations can improve the robustness or not and experimentally confirm our conclusions that more "implicit" leads to more robust performance.

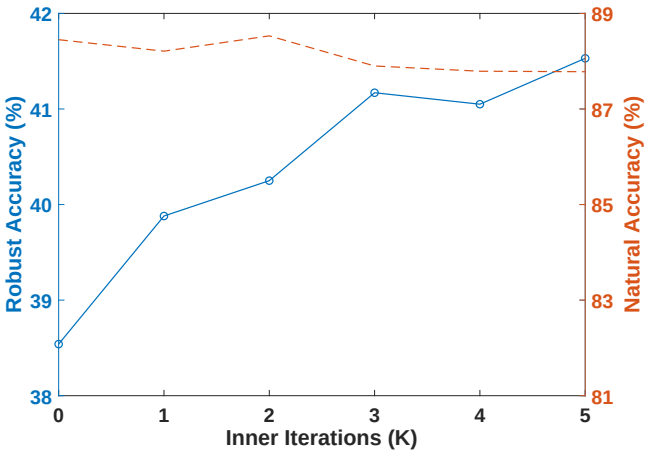

Figure 6. The robust accuracy and natural accuracy of IE-ResNet models with different inner iterations under the PGD-20 attack. The model is trained by PGD-3 with $\delta = 8/255$ and $\alpha = 2/255$.

The base model here is a Pre-Act ResNet-18 with initial channels equaling 64. From Figure 6, one can see that the inner iteration (K) does not have much impact on the natural accuracy while the robust accuracy increases sharply with the increment of the inner iteration (K). The empirical results also confirm our theory that more "implicit" ResNet may lead to more robust performance with respect to adversarial attacks. However, more inner iterations will be more computational consuming. Therefore, we leave finding a much more "implicit" architecture with less computational consumption as our future work.

6.2. Comparing with the Share-weight ResNets

Our IE-ResNet enhances the robustness of the vanilla ResNet by involving a single-step optimization in the original residual block, which increases the forward depth of the network. However, as shown in Table 6, simply increasing the forward depth for ResNet by sharing weights of two adjacent residual blocks with the same dimension (SwResNet-58) cannot improve the robustness as our IE-ResNet does. (Details on swResNet are shown in the supplementary.) On this account, the improvements of our model benefit from the architecture's inherent robustness as we have analysed in Section 4.2 rather than the computational increments.

Models	C&W (%)	PGD-20 (%)
ResNet-34	49.70	35.97
SwResNet-58	50.25	36.06
IE-ResNet-34	**51.17**	**37.13**

Table 6. Robust Accuracies for different models trained by PGD-3 on CIFAR-10 via C&W and PGD-20 attacks.

7. Conclusions

Although the training methods to improve the robustness of the neural network have been widely explored, few works studied the inherent robustness of the neural network. With the consideration that the vanilla ResNet is a kind of explicit Euler discretization, for the first time we explore the relationship between the adversarial robustness and the stability of the dynamic systems. From such an aspect, we analyze the stability of the vanilla ResNet and point out the reasons why ResNets are vulnerable to adversarial attacks. Furthermore, we propose a new architecture called IE-Skips to replace the original skip connections for the Residual Network family, inspired by the implicit Euler discretization method. Then we analyse the stability merits of IE-ResNet from the dynamic system view. In the end, we conduct various experiments to demonstrate that our IE-ResNet is more robust than the vanilla ResNet, which also validate our theoretical analysis. Our perspective of dynamic system stability for the neural networks robustness and Implicit Euler architectures may also be used in other neural neworks other than the Residual Network family.

References

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. IEEE, 2017.

Chen, T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. In *Advances in neural information processing systems*, pp. 6571–6583, 2018.

Cisse, M., Bojanowski, P., Grave, E., Dauphin, Y., and Usunier, N. Parseval networks: Improving robustness to adversarial examples. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 854–863. JMLR. org, 2017.

E, W. A proposal on machine learning via dynamical systems. *Communications in Mathematics and Statistics*, 5 (1):1–11, 2017.

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

Haber, E. and Ruthotto, L. Stable architectures for deep neural networks. *Inverse Problems*, 34(1):014004, 2017.

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*, pp. 770–778, 2016a.

He, K., Zhang, X., Ren, S., and Sun, J. Identity mappings in deep residual networks. In *European conference on computer vision*, pp. 630–645. Springer, 2016b.

Jakubovitz, D. and Giryes, R. Improving dnn robustness to adversarial attacks using jacobian regularization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 514–529, 2018.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

LeCun, Y. and Cortes, C. MNIST handwritten digit database. 2010. URL http://yann.lecun.com/exdb/mnist/.

Li, Q. and Hao, S. An optimal control approach to deep learning and applications to discrete-weight neural networks. *arXiv preprint arXiv:1803.01299*, 2018.

Lin, J., Gan, C., and Han, S. Defensive quantization: When efficiency meets robustness. *arXiv preprint arXiv:1904.08444*, 2019.

Lu, Y., Zhong, A., Li, Q., and Dong, B. Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations. *arXiv preprint arXiv:1710.10121*, 2017.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Moosavi-Dezfooli, S.-M., Fawzi, A., and Frossard, P. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2574–2582, 2016.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pp. 8024–8035, 2019.

Shafahi, A., Najibi, M., Ghiasi, M. A., Xu, Z., Dickerson, J., Studer, C., Davis, L. S., Taylor, G., and Goldstein, T. Adversarial training for free! In *Advances in Neural Information Processing Systems*, pp. 3353–3364, 2019.

Sun, K., Zhu, Z., and Lin, Z. Enhancing the robustness of deep neural networks by boundary conditional gan. *arXiv preprint arXiv:1902.11029*, 2019.

Svoboda, J., Masci, J., Monti, F., Bronstein, M. M., and Guibas, L. Peernets: Exploiting peer wisdom against adversarial attacks. *arXiv preprint arXiv:1806.00088*, 2018.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Wang, B., Yuan, B., Shi, Z., and Osher, S. J. Enresnet: Resnet ensemble via the feynman-kac formalism. *arXiv preprint arXiv:1811.10745*, 2018.

Xie, C., Wu, Y., Maaten, L. v. d., Yuille, A. L., and He, K. Feature denoising for improving adversarial robustness. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 501–509, 2019.

Yang, Y., Wu, J., Li, H., Li, X., Shen, T., and Lin, Z. Dynamical system inspired adaptive time stepping controller for residual network families. *arXiv preprint arXiv:1911.10305*, 2019.

Ye, N. and Zhu, Z. Bayesian adversarial learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 6892–6901. Curran Associates Inc., 2018.

Zhang, D., Zhang, T., Lu, Y., Zhu, Z., and Dong, B. You only propagate once: Painless adversarial training using maximal principle. *arXiv preprint arXiv:1905.00877*, 2019a.

Zhang, H., Yu, Y., Jiao, J., Xing, E. P., Ghaoui, L. E., and Jordan, M. I. Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573*, 2019b.

Zhang, J., Han, B., Wynter, L., Low, K. H., and Kankanhalli, M. Towards robust resnet: A small step but a giant leap. *arXiv preprint arXiv:1902.10887*, 2019c.

Zhu, M., Chang, B., and Fu, C. Convolutional neural networks combined with runge-kutta methods. *arXiv preprint arXiv:1802.08831*, 2018.

Zügner, D., Akbarnejad, A., and Günnemann, S. Adversarial attacks on neural networks for graph data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2847–2856, 2018.