Lyapunov-Stable Deep Equilibrium Models

Haoyu Chu[1,2,3], **Shikui Wei**[*1,3], **Ting Liu**[4], **Yao Zhao**[1,3], **Yuto Miyatake**[5]
[1]Institute of Information Science, Beijing Jiaotong University
[2]Graduate School of Information Science and Technology, Osaka University
[3]Beijing Key Laboratory of Advanced Information Science and Network Technology
[4]School of Computer Science, Northwestern Polytechnical University
[5]Cybermedia Center, Osaka University
19112001@bjtu.edu.cn, shkwei@bjtu.edu.cn, liuting@nwpu.edu.cn, yzhao@bjtu.edu.cn, miyatake@cas.cmc.osaka-u.ac.jp

Abstract

Deep equilibrium (DEQ) models have emerged as a promising class of implicit layer models, which abandon traditional depth by solving for the fixed points of a single nonlinear layer. Despite their success, the stability of the fixed points for these models remains poorly understood. By considering DEQ models as nonlinear dynamic systems, we propose a robust DEQ model named LyaDEQ with guaranteed provable stability via Lyapunov theory. The crux of our method is ensuring the Lyapunov stability of the DEQ model's fixed points, which enables the proposed model to resist minor initial perturbations. To avoid poor adversarial defense due to Lyapunov-stable fixed points being located near each other, we orthogonalize the layers after the Lyapunov stability module to separate different fixed points. We evaluate LyaDEQ models under well-known adversarial attacks, and experimental results demonstrate significant improvement in robustness. Furthermore, we show that the LyaDEQ model can be combined with other defense methods, such as adversarial training, to achieve even better adversarial robustness.

Introduction

Deep equilibrium models have demonstrated remarkable progress in various deep learning tasks, such as language modeling, image classification, semantic segmentation, compressive imaging, and optical flow estimation (Bai, Koltun, and Kolter 2020; Winston and Kolter 2020; Zhao, Zheng, and Yuan 2023; Bai et al. 2022). Unlike conventional neural networks that rely on stacking layers, DEQ models define their outputs as solutions to an input-dependent fixed points equation and use arbitrary black-box solvers to reach the fixed points without storing intermediate activations. As a result, DEQ models are categorized as implicit networks, presenting a unique approach to deep learning.

However, the robustness of DEQ models remains largely unexplored. As widely known, deep neural networks (DNNs) are susceptible to adversarial examples, which are crafted with minor perturbations to input images. Given the pervasive use of deep learning in various aspects of daily life, the emergence of adversarial examples poses a severe threat to the security of deep learning systems (Szegedy et al. 2013; Lu et al. 2023; Zhang et al. 2023). Hence, it is imperative to investigate the robustness of DEQ models. An intriguing question arises: can adversarial examples easily deceive DEQ models as well? If so, can we fundamentally mitigate this issue?

Wei and Kolter (2021) showed that DEQ models are also vulnerable to adversarial examples and considered ℓ_∞ certified robustness for DEQ models. They presented IBP-MonDEQ, a modification of monotone deep equilibrium layers that allows for the computation of lower and upper bounds on its output via interval bound propagation. Nevertheless, our experimental analysis revealed that IBP-MonDEQ does not provide significant improvement in adversarial robustness for some complex image recognition tasks. Likewise, Li, Wang, and Lin (2022) proposed a defense method for DEQ models based on certified training.

The deep learning community has shown a great interest in improving the adversarial robustness of neural networks. For another kind of implicit network, neural ordinary differential equations (Neural ODEs) (Chen et al. 2018), defense methods based on the Lyapunov method have emerged owing to their connection with dynamical systems. Kang et al. (2021) proposed a stable Neural ODE with Lyapunov-stable equilibrium points for defending against adversarial attacks. Rodriguez, Ames, and Yue (2022) proposed a method for training ODEs by using a control-theoretic Lyapunov condition for stability, leading to improvement on adversarial robustness.

Typically, a stable dynamical system implies that all solutions in some region around an equilibrium point (i.e., in a neighborhood of an equilibrium point) flow to that point. Lyapunov's direct method generalizes this concept by reasoning about convergence to states that minimize a potential Lyapunov function. Because Lyapunov theory deals with the effect of the initial perturbations on dynamic systems, integrating Lyapunov theory into implicit layers can automatically confer many benefits, such as adversarial robustness.

In this paper, we present a novel approach for improving the robustness of DEQ models through provable stability guaranteed by the Lyapunov theory. Unlike existing methods that rely on certified training or adversarial training (Wei and Kolter 2021; Li, Wang, and Lin 2022; Yang, Pang, and Liu 2022; Yang et al. 2023), our approach treats the DEQ model as a nonlinear dynamic system and ensures

*Corresponding author.

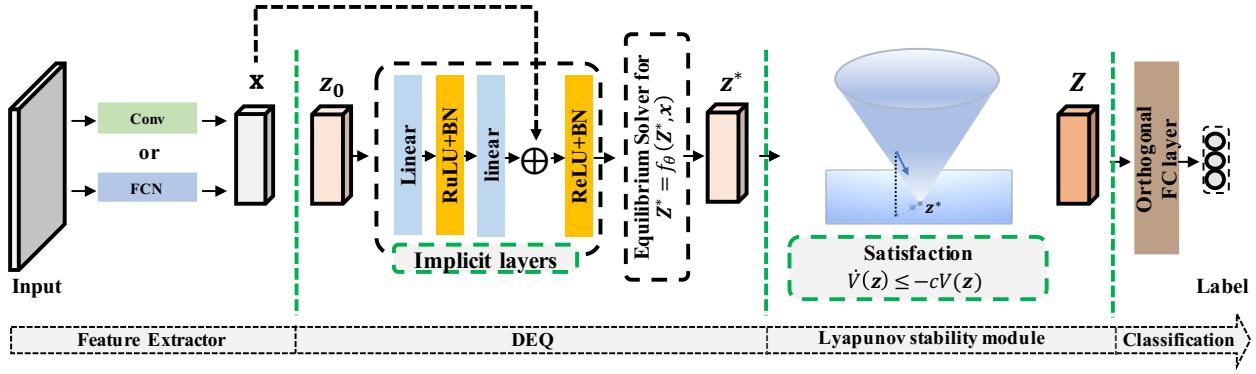


Figure 1: The scratch of the architecture of the LyaDEQ model. The blue arrow represents a state that locally satisfies the Lyapunov exponential stability condition.

that its fixed points are Lyapunov-stable, thereby keeping the perturbed fixed point within the same stable neighborhood as the unperturbed point and preventing successful adversarial attacks. Specially, we ensure the robustness of the DEQ model by jointly learning a convex positive definite Lyapunov function along with dynamics constrained to be stable according to these dynamics everywhere in the state space. Consequently, the minor adversarial perturbations added to the input image will hardly change the output of the DEQ model. Besides, for classification problems, Lyapunov-stable fixed points for different classes may be located near each other, leading to each stable neighborhood being very small, resulting in poor robustness against adversarial examples. To address this issue, we propose to use orthogonalization techniques to increase the distance between Lyapunov stable equilibrium points. The architecture of the LyaDEQ model is shown in Figure 1.

We name our proposed model LyaDEQ. Our main contributions are summarized as follows:

(1) We introduce Lyapunov stability theory into the DEQ model by considering it as a nonlinear system, which enables us to certify the stability of the fixed points. To the best of our knowledge, this is the first attempt to utilize the Lyapunov stability framework in the DEQ model.

(2) To address the poor adversarial defense caused by the small stable neighborhood of the fixed points, we introduce an orthogonal parametrized fully connected (FC) layer after the Lyapunov stability module to separate different Lyapunov-stable fixed points.

(3) The experimental results on MNIST, Street View House Numbers (SVHN), and CIFAR10/100 datasets demonstrate that the proposed LyaDEQ model consistently outperforms the baseline model in terms of robustness against adversarial attacks. These results validate the applicability of the Lyapunov theory to match the DEQ model and support the correctness of our theoretical analysis.

(4) We show that the LyaDEQ model can be combined with other adversarial training methods such as TRADES (Zhang et al. 2019), robust dataset (Ilyas et al. 2019), and PGD-AT (Madry et al. 2017), to achieve even better adversarial robustness.

Related Works

This section reviews works related to deep equilibrium models and Lyapunov theory in deep learning.

Deep Equilibrium Models

Motivated by an observation that the hidden layers of many existing deep sequence models converge towards some fixed points, DEQ models (Bai, Kolter, and Koltun 2019) find these fixed points via the root-finding method. Due to DEQ models suffering from unstable convergence to a solution and lacking guarantees that a solution exists, Winston and Kolter (2020) proposed a monotone operator equilibrium network, which guarantees stable convergence to a unique fixed point. Bai, Koltun, and Kolter (2020) proposed the multi-scale deep equilibrium model for handling large-scale vision tasks, such as ImageNet classification and semantic segmentation on high-resolution images. Later, they presented a regularization scheme for DEQ models that explicitly regularizes the Jacobian of the fixed-point update equations to stabilize the learning of equilibrium models (Bai, Koltun, and Kolter 2021b). Furthermore, Bai, Koltun, and Kolter (2021a) introduced neural deep equilibrium solvers for DEQ models to improve the speed/accuracy trade-off across diverse large-scale tasks. Li et al. (2021) proposed the multi-branch optimization-induced equilibrium models based on modeling the hidden objective function for the multi-resolution recognition task. Tsuchida et al. (2021) showed that solving a kernelized regularised maximum likelihood estimate as an inner problem in a deep declarative network yields a large class of DEQ architectures. Tsuchida and Ong (2023) presented a DEQ model that solves the problem of joint maximum a-posteriori estimation in a graphical model representing nonlinearly parameterized exponential family principal component analysis. Gilton, Ongie, and Willett (2021) presented an approach based on DEQ models for solving the linear inverse problems in imaging.

Lyapunov Theory in Deep Learning

Lyapunov functions are convenient tools for the stability certification of dynamical systems. Recently, many researchers have leveraged the Lyapunov stability theory to construct

provable, neural network-based safety certificates. Kolter and Manek (2019) used a learnable (i.e., defined by neural network architectures) Lyapunov function to modify a base dynamics model to ensure the stability of equilibrium. Richards, Berkenkamp, and Krause (2018) constructed a neural network Lyapunov function and a training algorithm to adapt them to the shape of the largest safe region for a closed-loop dynamical system. Chang et al. (2019) proposed to use anti-symmetric weight matrices to parametrize an RNN from the Lyapunov stability perspective, which enhances its long-term dependency.

Since the appearance of Neural ODEs, integrating Lyapunov methods into Neural ODEs has become a new trend. Inspired by LaSalle's theorem (an extension of Lyapunov's direct method), Takeishi and Kawahara (2021) proposed a deep dynamics model that can handle the stability of general types of invariant sets such as limit cycles and line attractors. They used augmented Neural ODEs (Dupont, Doucet, and Teh 2019) as the invertible feature transform for the provable existence of a stable invariant set. Massaroli et al. (2020) introduced stable neural flows whose trajectories evolve on monotonically non-increasing level sets of an energy functional parametrized by a neural network. Based on classical time-delay stability theory, Schlaginhaufen et al. (2021) proposed a new regularization term based on a neural network Lyapunov–Razumikhin function to stabilize neural delay differential equations.

Preliminaries

For a nonlinear system $\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u} = F(\boldsymbol{u})$, a state $\boldsymbol{u}^\star$ is a fixed point (or an equilibrium point) of a nonlinear system if $\boldsymbol{u}^\star$ satisfies $F(\boldsymbol{u}^\star) = \mathbf{0}$. A nonlinear system can have several (or infinitely many) isolated fixed points. One of the common interests in analyzing dynamical systems is the Lyapunov stability of the fixed points. A fixed point is stable means that the trajectories starting near $\boldsymbol{u}^\star$ remain around it all the time. More formally;

Definition 1 (Lyapunov stability)**.** An equilibrium $\boldsymbol{u}^\star$ is said to be stable in the sense of Lyapunov, if for every $\varepsilon > 0$, there exists $\delta > 0$ such that, if $\|\boldsymbol{u}(0) - \boldsymbol{u}^\star\| < \delta$, then $\|\boldsymbol{u}(t) - \boldsymbol{u}^\star\| < \varepsilon$ for all $t \geq 0$. If $\boldsymbol{u}^\star$ is stable, and $\lim_{t\to\infty} \|\boldsymbol{u}(t) - \boldsymbol{u}^\star\| = 0$, $\boldsymbol{u}^\star$ is said to be asymptotically stable. If $\boldsymbol{u}^\star$ is stable and for $\nu > 0$, if $\lim_{t\to\infty} \|\boldsymbol{u}(t) - \boldsymbol{u}^\star\| e^{\nu t} = 0$, $\boldsymbol{u}^\star$ is said to be exponentially stable.

Theorem 2 (Lyapunov stability theorem)**.** (Giesl and Hafstein 2015) Let $\boldsymbol{u}^\star$ be a fixed point. Let $V : \mathcal{U} \to \mathbb{R}$ be a continuously differentiable function, defined on a neighborhood $\mathcal{U}$ of $\boldsymbol{u}^\star$, which satisfies

(1) V has a minimum at $\boldsymbol{u}^\star$. A sufficient condition is $V(\boldsymbol{u}) \geq 0$ for all $\boldsymbol{u} \in \mathcal{U}$ and $V(\boldsymbol{u}) = 0 \Leftrightarrow \boldsymbol{u} = \boldsymbol{u}^\star$.

(2) V is strictly decreasing along solution trajectories in $\mathcal{U}$ except for the fixed point. A sufficient condition is $\dot{V}(\boldsymbol{u}) < 0$ for all $\boldsymbol{u} \in \mathcal{U}\backslash\{\boldsymbol{u}^\star\}$, where

$$\dot{V}(\boldsymbol{u}) = \frac{dV}{dt} = \nabla V(\boldsymbol{u})^T F(\boldsymbol{u}) < 0. \tag{1}$$

If such a function V exists, then it is called a Lyapunov function, and $\boldsymbol{u}^\star$ is *asymptotically stable*. Moreover, $\boldsymbol{u}^\star$ is *exponentially stable* if there exists positive definite V and some $K(\alpha) > 0$ such that

(3) $\|\boldsymbol{u}\|_2^2 \leq V(\boldsymbol{u}) \leq K(\alpha)\|\boldsymbol{u}\|_2^2$, for all $\boldsymbol{u}$ that $\|\boldsymbol{u}\| \leq \alpha$;

(4) $\dot{V}(\boldsymbol{u}) \leq -cV(\boldsymbol{u})$, $c > 0$.

Methodology

In this section, we first provide a dynamic system perspective for the DEQ model. Then, we introduce the Lyapunov stability framework, which is essential to our proposed model. Finally, we present the LyaDEQ model as a novel framework for enhancing the robustness of the DEQ model.

Considering a DEQ Model as a Nonlinear System

Given an input $\boldsymbol{x}$, a DEQ model (Bai, Kolter, and Koltun 2019) aims to specify a layer f_θ that finds the fixed points of the following iterative procedure

$$\boldsymbol{z}_{i+1} = f_\theta(\boldsymbol{z}_i, \boldsymbol{x}), \tag{2}$$

where $i = 0, \ldots, L-1$. Usually, we set $\boldsymbol{z}_0 = \mathbf{0}$ and choose the layer f_θ as a shallow block, such as a fully connected layer or convolutional layer.

Unlike a conventional neural network where the outputs are the activations from the L^{th} layer, the outputs of a DEQ model are the fixed points. One can alternatively find the fixed points $\boldsymbol{z}^\star = f_\theta(\boldsymbol{z}^\star, \boldsymbol{x})$ directly via root-finding algorithms rather than fixed points iteration alone:

$$f_\theta(\boldsymbol{z}^\star, \boldsymbol{x}) - \boldsymbol{z}^\star = 0. \tag{3}$$

Efficient root-finding algorithms, such as Broyden's method (Broyden 1965) and Anderson acceleration (Anderson 1965), can be applied to find this solution.

By defining $F(\boldsymbol{z}^\star) = f_\theta(\boldsymbol{z}^\star, \boldsymbol{x}) - \boldsymbol{z}^\star = 0$, we consider a DEQ model as a nonlinear dynamical system, i.e., a nonlinear system parameterized by a DEQ model. By doing so, we can use Lyapunov's theory to stabilize the fixed points and enable the DEQ model to resist minor initial perturbations on the inputs.

Lyapunov Stability Framework

Lyapunov's direct method is a powerful tool for studying the stability of dynamical systems. It aims to determine whether a system's final state, influenced by initial perturbations, can return to its original equilibrium state. Asymptotic stability, as defined in Definition 1, means that any initial state near the equilibrium state will eventually approach the equilibrium state. Exponential stability, on the other hand, ensures that the system's trajectory decays at a minimum attenuation rate.

In this paper, we consider the fixed points of the DEQ model as the equilibrium state and mainly focus on the adversarial perturbations added to the input images. We aim to ensure the stability of the deep equilibrium models by jointly learning a convex Lyapunov function along with dynamics constrained to be stable according to these dynamics everywhere in the state space. As a result, our proposed LyaDEQ model is anticipated to exhibit robustness against adversarial examples.

We use neural networks to learn a positive definite Lyapunov function V that satisfies conditions (1) and (3) in Theorem 2 and project outputs of a base dynamics model onto a space where condition (4) also holds (Kolter and Manek 2019).

Let $F(\boldsymbol{z}^\star) : \mathbb{R}^n \to \mathbb{R}^n$ be a basic dynamic system parametrized by a DEQ model, let $V : \mathbb{R}^n \to \mathbb{R}$ be a positive definite function, and c be a nonnegative constant, the Lyapunov-stable nonlinear dynamic model is defined as

$$\begin{aligned}\hat{F}(\boldsymbol{z}^\star) &= \text{Proj}(F(\boldsymbol{z}^\star), \{F : \nabla V(\boldsymbol{z}^\star)^T F \leq -cV(\boldsymbol{z}^\star)\}) \\ &= \begin{cases} F(\boldsymbol{z}^\star) & \text{if } \phi(\boldsymbol{z}^\star) \leq 0, \\ F(\boldsymbol{z}^\star) - \nabla V(\boldsymbol{z}^\star) \frac{\phi(\boldsymbol{z}^\star)}{\|\nabla V(\boldsymbol{z}^\star)\|_2^2} & \text{otherwise}, \end{cases}\end{aligned} \tag{4}$$

where $\phi(\boldsymbol{z}^\star) = \nabla V(\boldsymbol{z}^\star)^T F(\boldsymbol{z}^\star) + cV(\boldsymbol{z}^\star)$.

The Lyapunov function V is defined as positive definite and continuously differentiable, and has no local minima:

$$V(\boldsymbol{z}^\star) = \sigma_{k+1}(g(\boldsymbol{z}^\star) - g(0)) + \|\boldsymbol{z}^\star\|_2^2, \tag{5}$$

where σ_k is a positive convex non-decreasing function with $\sigma_k(0) = 0$, and g is represent as an input-convex neural network (ICNN) (Amos, Xu, and Kolter 2017):

$$\begin{aligned}\boldsymbol{q}_1 &= \sigma_0\left(W_0^I \boldsymbol{z}^\star + b_0\right), \\ \boldsymbol{q}_{i+1} &= \sigma_i\left(U_i \boldsymbol{q}_i + W_i^I \boldsymbol{z}^\star + b_i\right), i = 1, \ldots, k-1, \\ g(\boldsymbol{z}^\star) &\equiv \boldsymbol{q}_k,\end{aligned} \tag{6}$$

where W_i^I are real-valued weights, b_i are real-valued biases, and U_i are positive weights.

Proposition 1 The function V is convex in $\boldsymbol{z}^\star$ provided that all U_i are non-negative, and all functions σ_i are convex and non-decreasing.

Proof. The proof follows from the fact that non-negative sums of convex functions are convex and that the composition of a convex and convex non-decreasing function is also convex. For more detailed proof, please refer to (Boyd, Boyd, and Vandenberghe 2004). □

Through the procedures described above, we can therefore ensure that the DEQ model satisfies the conditions of the Lyapunov stability theorem. As a result, the fixed points of the modified DEQ model become exponentially stable.

LyaDEQ Model

As shown in Figure 1, our proposed model, LyaDEQ, consists of a feature extractor, a DEQ model, a Lyapunov stability module, and an orthogonal FC layer.

Feature extractor The feature extractor plays the role of dimensionality reduction. In our experiment, we choose a fully connected network (FCN) or ResNet (He et al. 2016) as the backbone of LyaDEQ.

DEQ A DEQ model ultimately finds the fixed points of a single function $\boldsymbol{z}^\star = f_\theta(\boldsymbol{z}^\star, \boldsymbol{x})$. We define the implicit layer f_θ as a feed-forward neural network, which can be written formally as

$$\begin{aligned}\boldsymbol{y} &= W_2^D \text{BN}(\text{ReLU}(W_1^D \boldsymbol{z}_0 + b) + b \\ f_\theta(\boldsymbol{z}, \boldsymbol{x}) &= \text{BN}(\text{ReLU}(\boldsymbol{x} + \boldsymbol{y})),\end{aligned} \tag{7}$$

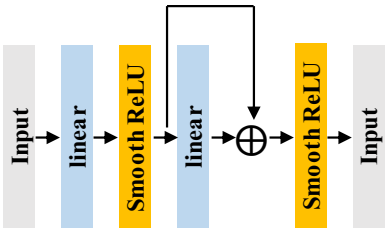


Figure 2: The architecture of our used ICNN.

where W^D are real-valued weights and BN represents the batch normalization operator.

We use Anderson acceleration (Anderson 1965) to find the fixed points of the DEQ model.

Lyapunov stability module We define ICNN as a 2-layer fully connected neural network. The network architecture is shown in Figure 2. The activation function σ is chosen as a smooth ReLU function:

$$\sigma(x) = \begin{cases} 0 & \text{if } x \leq 0, \\ x^2/2d & \text{if } 0 < x < d, \\ x - d/2 & \text{otherwise.} \end{cases} \tag{8}$$

Orthogonal FC layer As one can see from Definition 1, Lyapunov stability is established within a smaller stable neighborhood. For classification problems, Lyapunov-stable fixed points for different classes may be very close to each other, leading to each stable neighborhood may be very small, resulting in poor robustness against adversarial examples (see the experimental results on MNIST in Table 1).

We add an orthogonal FC layer after the Lyapunov stability module to increase the distance between Lyapunov stable equilibrium points. Given the output of the Lyapunov stability module $\boldsymbol{Z}$, the orthogonal FC layer will return the parametrized version $\boldsymbol{Z}$ so that $\boldsymbol{Z}^T \boldsymbol{Z} = \boldsymbol{I}$. The t-SNE visualization of the features after the orthogonal FC layer is shown in Figure 3.

Experiments

In this section, we first present the experimental setup. We then proceed to evaluate the performance of the LyaDEQ model against two white-box adversarial attacks. Subsequently, we demonstrate the effectiveness of the LyaDEQ model trained with adversarial training, comparing it to conventional convolutional neural networks (CNNs) in terms of robustness. Lastly, we conduct an ablation study to investigate the role of the orthogonal FC layer.

Setup

We conduct a set of experiments on three standard datasets MNIST (LeCun et al. 1998), CIFAR10/100 (Krizhevsky, Hinton et al. 2009), and SVHN (Netzer et al. 2011)

Benchmark	Model	Clean	Attack	$\epsilon = 2/255$	$\epsilon = 4/255$	$\epsilon = 6/255$	$\epsilon = 8/255$
MNIST	DEQ (baseline)	96.99	I-FGSM	50.24	49.89	49.67	49.34
			PGD	45.98	45.80	45.47	45.23
	DEQ w/ orthog. FC(ablation)	96.56	I-FGSM	29.98	29.91	29.60	29.44
			PGD	29.94	29.94	29.71	29.53
	IBP-MonDEQ	99.29	I-FGSM	**96.52**	**96.38**	**96.32**	**96.27**
			PGD	**94.87**	**94.80**	**94.69**	**94.57**
	LyaDEQ w/o orthog. FC (ours)	97.10	I-FGSM	23.18	23.28	23.21	23.25
			PGD	23.17	23.26	23.24	23.23
	LyaDEQ w/ orthog. FC (ours)	96.59	I-FGSM	50.78 (+0.54)	50.60 (+0.71)	50.55 (+0.88)	50.43 (+1.09)
			PGD	50.72 (+4.74)	50.54 (+4.74)	50.52 (+5.05)	50.35 (+5.12)
SVHN	DEQ (baseline)	95.37	I-FGSM	68.09	61.78	56.77	51.65
			PGD	67.06	60.89	56.38	51.22
	DEQ w/ orthog. FC (ablation)	95.63	I-FGSM	67.30	59.96	54.27	48.79
			PGD	66.74	60.31	55.59	50.02
	IBP-MonDEQ	91.06	I-FGSM	57.71	55.71	53.78	51.58
			PGD	57.75	55.74	54.15	52.04
	LyaDEQ w/o orthog. FC (ours)	95.28	I-FGSM	**72.16 (+4.07)**	**72.08 (+10.30)**	**71.84 (+15.07)**	**71.42 (+19.77)**
			PGD	68.95	67.88	67.11	65.60
	LyaDEQ w/ orthog. FC (ours)	95.21	I-FGSM	69.41	69.01	68.55	67.81
			PGD	**71.35 (+4.29)**	**71.23 (+10.34)**	**70.95 (+14.57)**	**70.47 (+19.25)**
CIFAR10	DEQ (baseline)	87.71	I-FGSM	34.25	23.00	16.65	12.36
			PGD	33.37	22.87	16.81	11.12
	DEQ w/ orthog. FC (ablation)	87.62	I-FGSM	35.46	23.48	16.91	11.61
			PGD	32.67	22.07	15.92	10.89
	IBP-MonDEQ	80.25	I-FGSM	25.87	24.54	23.59	22.24
			PGD	27.59	26.21	25.04	23.58
	LyaDEQ w/o orthog. FC (ours)	87.80	I-FGSM	43.80	43.99	43.75	43.45
			PGD	**47.09 (+13.72)**	**46.94 (+24.07)**	**46.70 (+29.89)**	**46.12 (+35.00)**
	LyaDEQ w/ orthog. FC (ours)	87.87	I-FGSM	**47.38 (+13.13)**	**47.31 (+24.31)**	**46.71 (+30.06)**	**45.80 (+33.44)**
			PGD	45.04	44.81	44.50	43.96
CIFAR100	DEQ (baseline)	61.23	I-FGSM	15.62	8.59	5.94	4.15
			PGD	11.85	6.71	4.47	3.10
	DEQ w/ orthog. FC (ablation)	61.41	I-FGSM	14.75	8.13	5.59	3.85
			PGD	13.81	7.70	5.16	3.39
	IBP-MonDEQ	44.78	I-FGSM	9.15	8.32	7.68	7.04
			PGD	8.88	8.21	7.56	7.04
	LyaDEQ w/o orthog. FC (ours)	60.52	I-FGSM	22.20	22.16	21.77	21.07
			PGD	20.25	19.78	19.10	18.18
	LyaDEQ w/ orthog. FC (ours)	61.30	I-FGSM	**23.82 (+8.20)**	**23.76 (+15.17)**	**23.73 (+17.79)**	**23.35 (+19.20)**
			PGD	**21.86 (+10.01)**	**21.53 (+14.82)**	**20.65 (+16.18)**	**19.55 (+16.45)**

Table 1: Classification accuracy on MNIST, SVHN, and CIFAR. Results that surpass all competing methods are bold. The second best result is with the underline. The performance gain in parentheses is compared with the baseline model. The input is the test set of CIFAR10. We abbreviate the LyaDEQ model without the orthogonal FC layer as LyaDEQ w/o orthog. FC.

Training configurations We use PyTorch (Paszke et al. 2017) framework for the implementation. For MNIST, we use 1-layer FCN as the feature extractor to reduce the dimension from 28×28 to 64. For SVHN and CIFAR10, we use ResNet20 (He et al. 2016) to reduce the dimension from 32×32×3 to 64 (128 for CIFAR100).

For optimization, we use Adam algorithm (Kingma and Ba 2014) with betas=(0.9, 0.999). We set the initial learning rate to 0.001 and set the learning rate of each parameter group using a cosine annealing schedule. The training epochs for MNIST, SVHN, and CIFAR are set to 10, 40, and 50.

The configurations of adversarial attacks We test the performance of the original DEQ model and our proposed LyaDEQ model on two white-box adversarial attacks: iterative fast gradient sign method (I-FGSM) (Kurakin, Goodfellow, and Bengio 2018) and project gradient descent (PGD) (Madry et al. 2017).

I-FGSM As an iterative version based FGSM (Goodfellow, Shlens, and Szegedy 2014), I-FGSM computes an adversarial example by multiple gradients:

$$\boldsymbol{x}_{i+1}^{adv} = \mathrm{Clip}_{x,\epsilon}\{\boldsymbol{x}_i^{adv} + \alpha\,\mathrm{sign}(\nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}_i^{adv}, \boldsymbol{y}))\}, \quad (9)$$

where α is the step size, $\mathrm{Clip}_{x,\epsilon}$ means clipping perturbed images within $[x-\epsilon, x+\epsilon]$, and $\boldsymbol{x}_0^{adv} = \boldsymbol{x}$.

PGD attack is a universal attack utilizing the local first order information about the network:

$$\boldsymbol{x}_{i+1}^{adv} = \Pi_{\boldsymbol{x}+\mathcal{S}}(\boldsymbol{x}_i^{adv} + \alpha\,\mathrm{sign}(\nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}_i^{adv}, \boldsymbol{y}))), \quad (10)$$

where $\Pi_{\boldsymbol{x}+\mathcal{S}}$ represents the projection on ϵ-ball, $\mathcal{S} \subseteq \mathbb{R}^d$. A uniform random noise is first added to the clean image $\boldsymbol{x}$,

that is $\boldsymbol{x}_0^{adv} = \boldsymbol{x} + \mathcal{U}[-\epsilon, \epsilon]$. We set the size of perturbation ϵ of PGD in the infinite norm sense.

For both PGD and I-FGSM, the step size α is set to 1/255, and the number of steps n is calculated as $n = \lfloor \min(\epsilon \cdot 255 + 4, \epsilon \cdot 255 \cdot 1.25) \rfloor$.

The Robustness of LyaDEQ Model Against Adversarial Examples

Table 1 displays the results of our experiments regarding classification accuracy and robustness against adversarial examples. Regarding classification accuracy on clean data, our proposed model achieves comparable performance with the baseline model.

Regarding robustness against adversarial examples, we evaluate the effectiveness of our proposed LyaDEQ model in defending against white-box attacks with attack radii ranging from $\epsilon = 2/255$ to $\epsilon = 8/255$. Our experimental results demonstrate that the LyaDEQ model outperforms the baseline model on each dataset. For instance, compared with the DEQ model under PGD attack with $\epsilon = 8/255$, the LyaDEQ model exhibits a 5.12%, 19.25%, 32.48%, and 16.45% improvement on MNIST, SVHN, CIFAR10, and CIFAR100 datasets, respectively. These findings confirm that the Lyapunov stability module can significantly enhance the robustness of the DEQ model.

The results presented in Table 1 demonstrate that the magnitude of the accuracy boost under adversarial attack increases with increasing attack radii for all datasets. For instance, LyaDEQ model under I-FGSM with attack radii $\epsilon = 2/255$, $\epsilon = 4/255$, $\epsilon = 6/255$ and $\epsilon = 8/255$ has a 13.13%, 24.31%, 30.06%, 33.44% boost respectively on CIFAR10. This further corroborates the effectiveness of our proposed approach.

In addition, while the experimental results of the IBP-MonDEQ model on MNIST are much better than our method, it performs poorly on SVHN, CIFAR10, and CIFAR100 datasets. We consider that the reason for conflict lies in the complexity of image datasets. For MNIST, the scene of the image is quite simple. In contrast, for SVHN and CIFAR, the scene of the image is more complex. Thus, our proposed LyaDEQ model is better suited for handling complex image recognition tasks and can be widely utilized in commonly used datasets.

Explanations on performance improvement: The core concept of our proposed method revolves around utilizing the Lyapunov direct method to ensure the stability of the fixed points of the DEQ models. According to the Lyapunov stability theory, if the magnitude of perturbations on $\boldsymbol{z}$ exceeds the stable neighborhood, the impact on the final outcome becomes unpredictable, resulting in potential misclassification by the model. Conversely, if the magnitude of perturbations remains within the stable neighborhood, the final outcome remains unaffected, enabling our model to effectively resist adversarial noise and maintain robustness against adversarial attacks.

LyaDEQ Model With Adversarial Training

Our method is orthogonal to other adversarial defense methods, such as adversarial training, which means we can combine the LyaDEQ model with adversarial training to achieve further defense performance. We choose three commonly used adversarial training methods, TRADES (Zhang et al. 2019), robust dataset (RD) (Ilyas et al. 2019) and PGD-AT (PAT) (Madry et al. 2017).

Radius	Attack	+TRADES	+RD	+PAT
$\epsilon = 2/255$	I-FGSM	**72.81**	51.96	60.46
	PGD	48.48	52.00	**60.43**
$\epsilon = 4/255$	I-FGSM	**72.68**	51.78	60.34
	PGD	48.39	51.85	**60.42**
$\epsilon = 6/255$	I-FGSM	**72.66**	51.62	60.18
	PGD	48.35	51.54	**60.41**
$\epsilon = 8/255$	I-FGSM	**72.42**	51.51	60.08
	PGD	48.33	51.63	**60.23**

Table 2: Classification accuracy of the LyaDEQ model combined with adversarial training method on CIFAR10 under adversarial attacks.

Radius	Attack	ResNet56	VGG16	WResNet
$\epsilon = 2/255$	I-FGSM	56.79	58.63	54.41
	PGD	52.70	55.90	51.58
$\epsilon = 4/255$	I-FGSM	49.35	51.50	46.05
	PGD	45.07	49.24	43.74
$\epsilon = 6/255$	I-FGSM	44.88	45.60	41.43
	PGD	41.35	43.70	39.68
$\epsilon = 8/255$	I-FGSM	40.51	37.94	37.08
	PGD	36.58	36.46	35.11

Table 3: Classification accuracy of the conventional CNNs on CIFAR10 under adversarial attacks.

TRADES is a defense method to trade adversarial robustness off against accuracy via combining tricks of warmup, early stopping, weight decay, batch size, and other hyper-parameter settings. In our experiments, we set perturbation epsilon = 0.031, perturbation step size = 0.007, number of iterations = 10, beta = 6.0 on the training dataset.

RD is created by removing non-robust features from the dataset, which yields good robust accuracy on the unmodified test set.

PAT is a defense method to inject adversarial examples that generated by the PGD attack into training data. It is worth mentioning that Yang, Pang, and Liu (2022) also used PAT to train DEQ models. In our experiments, we set perturbation epsilon = 0.031, perturbation step size = 0.00784, and number of iterations = 7 on the training dataset.

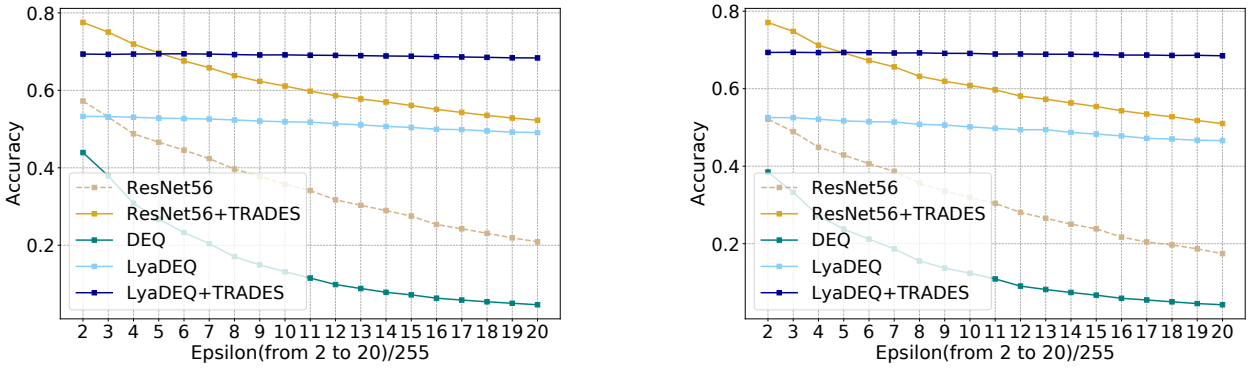
From Table 2, we see that training the LyaDEQ model with adversarial training methods can further improve robustness against adversarial examples. For example, training the LyaDEQ model with TRADES, the robust dataset and PAT show a 25.43%, 4.58%, and 13.08% boost respectively on CIFAR10 under I-FGSM attack with $\epsilon = 2/255$.

Comparison With the Conventional CNNs

We perform experiments to compare the robustness of our proposed LyaDEQ model with three conventional CNNs, ResNet56 (He et al. 2016), VGG16 (Simonyan and Zisser-

(a) LyaDEQ (clean) (b) LyaDEQ w/o orthog. FC (clean) (c) LyaDEQ (adv) (d) LyaDEQ w/o orthog. FC (adv)

Figure 3: t-SNE visualization results on the features after the orthogonal FC layer. 'adv' means test on the adversarial dataset.



(a) Under I-FGSM attacks (b) Under PGD attacks

Figure 4: Comparison between the DEQ model, the LyaDEQ model, and conventional CNNs on CIFAR10.

man 2014), and WideResNet (WResNet) (Zagoruyko and Komodakis 2016) (with depth = 28, widen factor = 10). Table 3 summarizes the experimental results. While the robustness of the LyaDEQ model is inferior to that of conventional CNNs when the value of ϵ is small (e.g., when $\epsilon = 2/255, 4/255$), the robustness of the LyaDEQ model is significantly better when the value of ϵ is getting larger (e.g., when $\epsilon = 6/255, 8/255$) than these neural networks. For instance, under the PGD attack with $\epsilon = 8/255$, the accuracy of the LyaDEQ model is 7.38%, 7.50%, and 8.85% higher than that of ResNet56, VGG16, and WResNet, respectively.

This is because the conventional CNNs are very sensitive to the radius of the adversarial attacks. Increasing the radius has a significant impact on these networks, when the magnitude is increased from $\epsilon = 2/255$ to $\epsilon = 8/255$, the accuracy of ResNet56, VGG16, and WideResNet under I-FGSM attacks decreases by 16.28%, 20.69%, and 17.33%, respectively. Whereas, the accuracy of the LyaDEQ model only decreases by 1.58% in the same case, which shows that the LyaDEQ model is insensitive to the radius of the adversarial attack. Its insensitivity to the radius of adversarial attack becomes more evident when the value of ϵ is much larger. Figure 4 provides a detailed comparison between DEQ, LyaDEQ, LyaDEQ trained by TRADES, ResNet56, and ResNet trained by TRADES under adversarial attacks ranging from $\epsilon = 2/255$ to $\epsilon = 20/255$, further validating the robustness of the LyaDEQ model.

An Ablation Study

As an ablation study, we test the LyaDEQ model without the orthogonal FC layer and the DEQ model with the orthogonal FC Layer. As we expected, adding the orthogonal FC layer to the DEQ model does not have a significant impact. From Table 1, we find that in most cases, the orthogonal FC layer indeed plays a part in improving the accuracy of the LyaDEQ model under adversarial attack. Especially, on the MNIST dataset, the absence of the orthogonal FC layer incurred a substantial 20% loss in accuracy, underscoring its crucial contribution to robustness.

In addition, we notice the orthogonal FC layer occasionally incurs a minor deleterious effect. Nevertheless, this negative impact is acceptable when juxtaposed against the overall robustness improvement brought by the LyaDEQ model.

Conclusions

Inspired by Lyapunov stability theory, we introduced a provably stable variant of DEQ models. Our proposed model consists of a feature extractor, a DEQ model, a Lyapunov stability module, and an orthogonal FC layer. The Lyapunov stability module ensures the fixed points of the DEQ model are Lyapunov stable, and the orthogonal FC layer separates different Lyapunov-stable fixed points. Our findings highlighted the proposed method in improving the robustness of the DEQ model.

Acknowledgments

This work was supported in part by the National Key R&D Program of China (No.2021ZD0112100), the National Natural Science Foundation of China (No.62106201, No.61972022, No.U1936212, No.62120106009), the China Scholarship Council (No.202207090082), JSPS KAKENHI (20H01822, 20H00581, 21K18301), and JST PRESTP (JPMJPR2129).

References

Amos, B.; Xu, L.; and Kolter, J. Z. 2017. Input convex neural networks. In *International Conference on Machine Learning*, 146–155. PMLR.

Anderson, D. G. 1965. Iterative procedures for nonlinear integral equations. *Journal of the ACM (JACM)*, 12(4): 547–560.

Bai, S.; Geng, Z.; Savani, Y.; and Kolter, J. Z. 2022. Deep Equilibrium Optical Flow Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 620–630.

Bai, S.; Kolter, J. Z.; and Koltun, V. 2019. Deep equilibrium models. *Advances in Neural Information Processing Systems*, 32.

Bai, S.; Koltun, V.; and Kolter, J. Z. 2020. Multiscale deep equilibrium models. *Advances in Neural Information Processing Systems*, 33: 5238–5250.

Bai, S.; Koltun, V.; and Kolter, J. Z. 2021a. Neural deep equilibrium solvers. In *International Conference on Learning Representations*.

Bai, S.; Koltun, V.; and Kolter, J. Z. 2021b. Stabilizing equilibrium models by jacobian regularization. *arXiv preprint arXiv:2106.14342*.

Boyd, S.; Boyd, S. P.; and Vandenberghe, L. 2004. *Convex optimization*. Cambridge university press.

Broyden, C. G. 1965. A class of methods for solving nonlinear simultaneous equations. *Mathematics of computation*, 19(92): 577–593.

Chang, B.; Chen, M.; Haber, E.; and Chi, E. H. 2019. AntisymmetricRNN: A dynamical system view on recurrent neural networks. *arXiv preprint arXiv:1902.09689*.

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

Dupont, E.; Doucet, A.; and Teh, Y. W. 2019. Augmented neural odes. *Advances in Neural Information Processing Systems*, 32.

Giesl, P.; and Hafstein, S. 2015. Review on computational methods for Lyapunov functions. *Discrete & Continuous Dynamical Systems-B*, 20(8): 2291.

Gilton, D.; Ongie, G.; and Willett, R. 2021. Deep equilibrium architectures for inverse problems in imaging. *IEEE Transactions on Computational Imaging*, 7: 1123–1133.

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

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Ilyas, A.; Santurkar, S.; Tsipras, D.; Engstrom, L.; Tran, B.; and Madry, A. 2019. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32.

Kang, Q.; Song, Y.; Ding, Q.; and Tay, W. P. 2021. Stable neural ode with lyapunov-stable equilibrium points for defending against adversarial attacks. *Advances in Neural Information Processing Systems*, 34: 14925–14937.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kolter, J. Z.; and Manek, G. 2019. Learning stable deep dynamics models. *Advances in neural information processing systems*, 32.

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

Kurakin, A.; Goodfellow, I. J.; and Bengio, S. 2018. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, 99–112. Chapman and Hall/CRC.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Li, M.; Wang, Y.; and Lin, Z. 2022. Cerdeq: Certifiable deep equilibrium model. In *International Conference on Machine Learning*, 12998–13013. PMLR.

Li, M.; Wang, Y.; Xie, X.; and Lin, Z. 2021. Optimization inspired Multi-Branch Equilibrium Models. In *International Conference on Learning Representations*.

Lu, S.; Wang, M.; Wang, D.; Wei, X.; Xiao, S.; Wang, Z.; Han, N.; and Wang, L. 2023. Black-box attacks against log anomaly detection with adversarial examples. *Information Sciences*, 619: 249–262.

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

Massaroli, S.; Poli, M.; Bin, M.; Park, J.; Yamashita, A.; and Asama, H. 2020. Stable neural flows. *arXiv preprint arXiv:2003.08063*.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading digits in natural images with unsupervised feature learning.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch.

Richards, S. M.; Berkenkamp, F.; and Krause, A. 2018. The lyapunov neural network: Adaptive stability certification for safe learning of dynamical systems. In *Conference on Robot Learning*, 466–476. PMLR.

Rodriguez, I. D. J.; Ames, A.; and Yue, Y. 2022. LyaNet: A Lyapunov framework for training neural ODEs. In *International Conference on Machine Learning*, 18687–18703. PMLR.

Schlaginhaufen, A.; Wenk, P.; Krause, A.; and Dorfler, F. 2021. Learning Stable Deep Dynamics Models for Partially Observed or Delayed Dynamical Systems. *Advances in Neural Information Processing Systems*, 34: 11870–11882.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

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

Takeishi, N.; and Kawahara, Y. 2021. Learning dynamics models with stable invariant sets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 9782–9790.

Tsuchida, R.; and Ong, C. S. 2023. Deep equilibrium models as estimators for continuous latent variables. In *International Conference on Artificial Intelligence and Statistics*, 1646–1671. PMLR.

Tsuchida, R.; Yong, S. Y.; Armin, M. A.; Petersson, L.; and Ong, C. S. 2021. Declarative nets that are equilibrium models. In *International Conference on Learning Representations*.

Wei, C.; and Kolter, J. Z. 2021. Certified robustness for deep equilibrium models via interval bound propagation. In *International Conference on Learning Representations*.

Winston, E.; and Kolter, J. Z. 2020. Monotone operator equilibrium networks. *Advances in neural information processing systems*, 33: 10718–10728.

Yang, Z.; Li, P.; Pang, T.; and Liu, Y. 2023. Improving Adversarial Robustness of Deep Equilibrium Models with Explicit Regulations Along the Neural Dynamics.

Yang, Z.; Pang, T.; and Liu, Y. 2022. A Closer Look at the Adversarial Robustness of Deep Equilibrium Models. *Advances in Neural Information Processing Systems*, 35: 10448–10461.

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

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E.; El Ghaoui, L.; and Jordan, M. 2019. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, 7472–7482. PMLR.

Zhang, Y.; Tan, Y.-a.; Sun, H.; Zhao, Y.; Zhang, Q.; and Li, Y. 2023. Improving the invisibility of adversarial examples with perceptually adaptive perturbation. *Information Sciences*.

Zhao, Y.; Zheng, S.; and Yuan, X. 2023. Deep Equilibrium Models for Snapshot Compressive Imaging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 3642–3650.