CerDEQ: Certifiable Deep Equilibrium Model

Mingjie Li, Yisen Wang, Zhouchen Lin
Peking University

Certifiable Robustness and DEQ

IBP Certified Robustness

Interval Bound Propagation (IBP) is the widely used effective way to obtain the relaxed output bounds for the input perturbations. The IBP bound is obtained by solving the following problems layer by layer.

$$\bar{x}_i := \max_{\|x'-x\|_p<\epsilon} \{e_i^\top y : y = \sigma(Wx' + b)\}$$

$$\underline{x}_i := \min_{\|x'-x\|_p<\epsilon} \{e_i^\top y : y = \sigma(Wx' + b)\}$$

Finally we can get a convex output set for classification, which contains the models real output for the worst perturbation. If we can correctly classified the relaxed set, then the model is certifiably robust on such samples. We can also use the output set to do the certified training to enhance models robustness.

DEQ models

Given input x, the output of the DEQ model is the solution of the following fixed point equation

$$z = \sigma(Wz + Ux + b)$$

which can be regarded as an weight-tied explicit model with infinite depth.

Certifiable Deep Equilibrium Models

Problems and Motivations

- The output bounds for deep explicit models are larger and makes certified training harder. Which lead to the descending performance with respect to the depth. While the performance of DEQ does not rely on the depth.
- Deep Equilibrium Models enjoys controllable global Lipschitz and its output bounds are better than explicit models as our following experiments shows.
- $\bar{z}_i \coloneqq \max_{\|x'-x\|_p<\epsilon} \{e_i^\top z : z = \sigma(Wz + Ux' + b)\}$ can not be directly obtained as in explicit models.

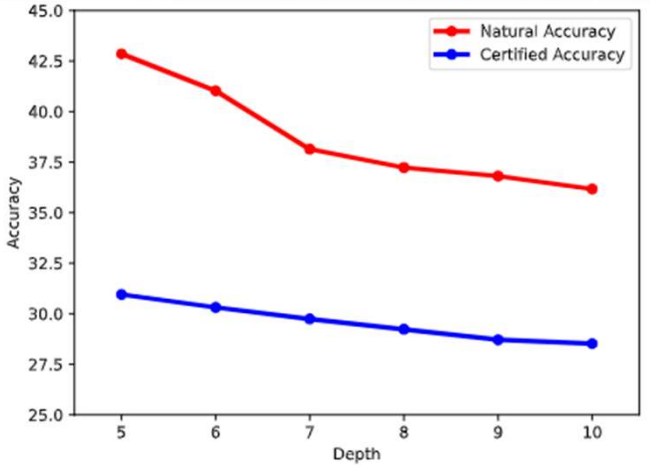


Figure 1. The certified error of CNNs with BN of different depth, the models are trained on CIFAR-10 and evaluated under $\epsilon = 8/255$.

Certifiable Deep Equilibrium Models

Use Adjoint DEQ to obtain the output bounds

For DEQ, we can obtain its output bounds with the Adjoint DEQ:

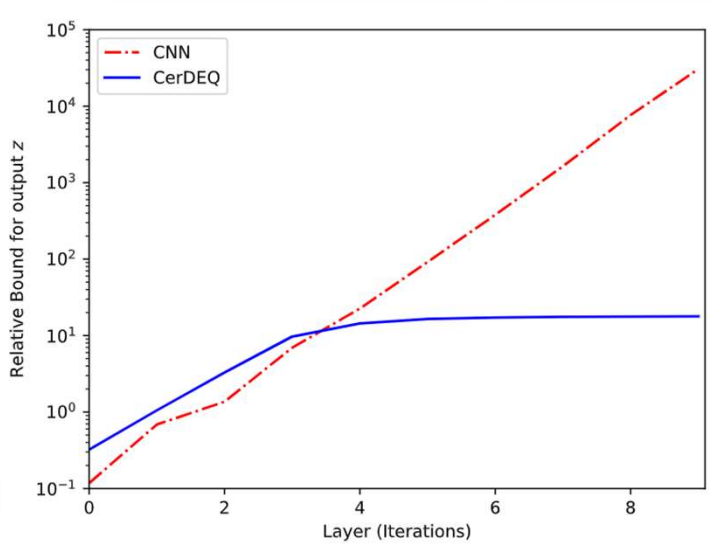
$$z^* = \sigma(Wz^* + Ux + b)$$

$$\begin{pmatrix} \overline{z^*} \\ \underline{z^*} \end{pmatrix} = \sigma\left(\begin{pmatrix} W_+ & W_- \\ W_- & W_+ \end{pmatrix}\begin{pmatrix} \overline{z^*} \\ \underline{z^*} \end{pmatrix} + \begin{pmatrix} \mathrm{U}_+\overline{x} + U_-\underline{x} \\ \mathrm{U}_-\overline{x} + U_+\underline{x} \end{pmatrix} + b\right)$$

The following equation needs to satisfy to ensure the convergence of DEQ and its adjoint one:

$$\left\|\begin{pmatrix} W_+ & W_- \\ W_- & W_+ \end{pmatrix}\right\|_2 < 1$$

One direct way to ensure the convergence is the weight normalization with scaling.



Certifiable Deep Equilibrium Models

北京大学 PEKING UNIVERSITY

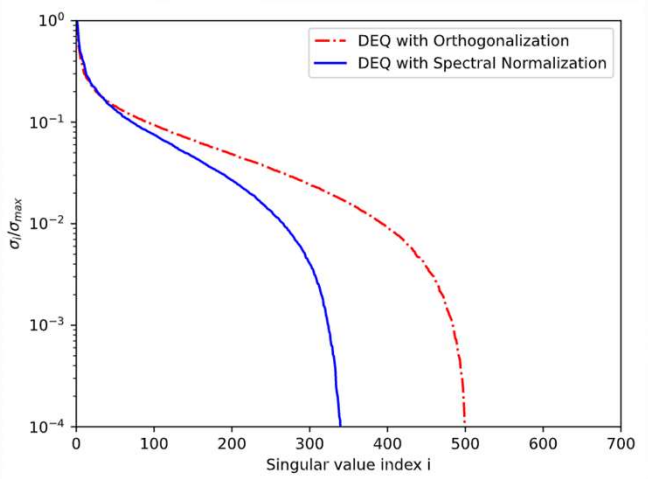
Weight Orthogonalization

As shown in the right, scaling the weights with large numbers may lead the weight matrix to be low rank. In other words, the width of DEQ may become narrow in practice.

To alleviate the problem, we use Bjorck Orthogonalization to project the weights after each update to its nearest orthogonal matrix and then do the scaling:

$$A_{k+1} = \frac{15}{8}A_k - \frac{5}{4}A_k\left(A_k^\top A_k\right) + \frac{3}{8}A_k(A_k^\top A_k)(A_k^\top A_k)$$

As shown in the right, our method can alleviate the phenomena and lead to better performance.



Model	Standard Error	Certified Error
DEQ+WN	$56.34 \pm 0.32\%$	$69.84 \pm 0.13\%$
DEQ+SN	$57.43 \pm 0.41\%$	$68.66 \pm 0.15\%$
CerDEQ	$\mathbf{53.43 \pm 0.33\%}$	$\mathbf{67.21 \pm 0.12\%}$

Certifiable Deep Equilibrium Models

CerDEQ's weight initialization

- Former initialization for explicit models is not suitable for our DEQ model. Therefore, we need to design our initialization methods. Since W is orthogonalized, we only need to consider the initialization of U.

- If we use gaussian distribution to initialize U, i.e., $U \sim \mathcal{N}(0, \sigma)$ and W is row-orthogonal with $\|W\|_2 \leq \frac{1}{\sqrt{n}}$, with $U \in \mathbb{R}^{m \times n_u}$ and $W \in \mathbb{R}^{m \times n}$. We can obey the following relationship for the DEQ layers input bound Δ_{in} and output bound Δ_{out}.

$$\frac{\mathbb{E}[\Delta_{out}]}{\mathbb{E}[\Delta_{in}]} \leq \frac{n_u \mathbb{E}[|U|]}{2 - \sqrt{n}\|W\|_2} \leq n_u \mathbb{E}[|U|]$$

Certifiable Deep Equilibrium Models

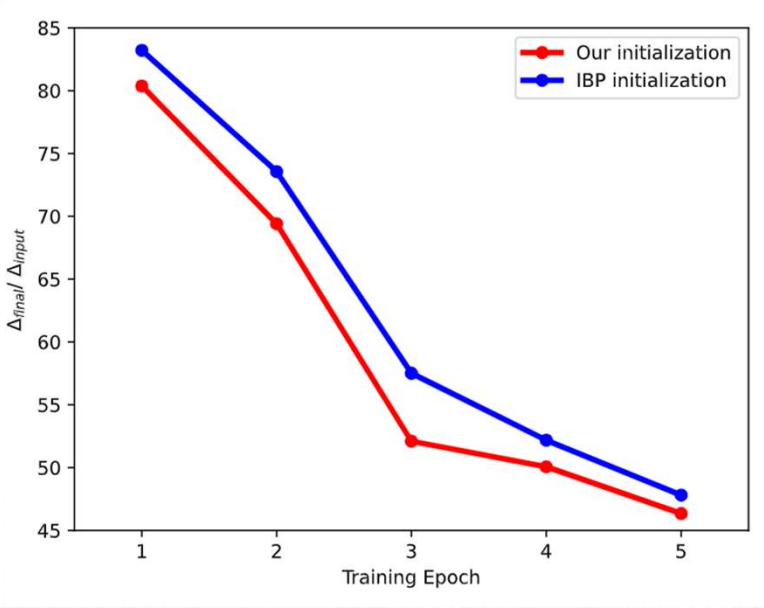
CerDEQ's weight initialization

If the elements of U obeys the Gaussian distribution:

$$\mathbb{E}[|U|] = \sqrt{\frac{2}{\pi}}$$

From former proposition, one can see that if we need to led the output bounds controllable in the beginning, i.e., let $\frac{\mathrm{E}[\Delta_{out}]}{\mathrm{E}[\Delta_{in}]} \leq 1$, we need to set:

$$\sigma = \frac{\sqrt{\pi}}{\sqrt{2}n_u}$$



CerDEQ — Results on CIFAR-10

北京大学 PEKING UNIVERSITY

70 epochs certified training with $\epsilon = 8/255$

Model	Standard Error	Certified Error
CNN-7	$56.64 \pm 0.48\%$	$68.81 \pm 0.24\%$
WideResNet	$56.74 \pm 0.40\%$	$68.79 \pm 0.29\%$
ResNeXt	$59.33 \pm 0.40\%$	$70.62 \pm 0.59\%$
CerDEQ (ours)	$\mathbf{53.43 \pm 0.33\%}$	$\mathbf{67.21 \pm 0.12\%}$

200 epochs certified training with $\epsilon = 8/255$

Model	Standard Error	Certified Error
CNN-7-BN	$51.72 \pm 0.40\%$	$65.58 \pm 0.24\%$
WideResNet	$51.95 \pm 0.32\%$	$65.91 \pm 0.14\%$
ResNeXt	$53.68 \pm 0.33\%$	$66.91 \pm 0.40\%$
CerDEQ (ours)	$\mathbf{50.34 \pm 0.33\%}$	$\mathbf{64.98 \pm 0.26\%}$
CerDEQ (best)	$\mathbf{49.97\%}$	$\mathbf{64.72\%}$

CerDEQ — Results on Tiny-ImageNet

北京大学 PEKING UNIVERSITY

Empirical Results after certified training

- $\epsilon = \frac{1}{255}$:

Model	Standard Error	Certified Error
CNN-7	74.29%	82.36%
WideResNet	74.59%	82.75%
ResNeXt	78.91%	85.78%
ℓ_∞-dist Net	78.18%	83.69%
CerDEQ	**73.51%**	**82.16%**

- $\epsilon = \frac{8}{255}$:

Model	Standard Error	Certified Error
CNN-7 (Crown-IBP)	90.76%	95.98%
CNN-7 (Fast-IBP)	89.69%	95.44%
ℓ_∞-dist Net	88.99%	**94.22%**
CerDEQ (ours)	**87.98%**	94.45%

Thanks for Watching!