Zhen LIANG, Taoran WU, Wanwei LIU, Bai XUE, Wenjing YANG, Ji WANG, Zhengbin PANG, 2023. Towards robust neural networks via a global and monotonically decreasing robustness training strategy. *Frontiers of Information Technology & Electronic Engineering*, 24(10):1375-1389.
https://doi.org/10.1631/FITEE.2300059

Towards robust neural networks via a global and monotonically decreasing robustness training strategy

Key words: Robust neural networks; Training method; Drawdown risk; Global robustness training; Monotonically decreasing robustness

Corresponding author: Wanwei LIU
E-mail: wwliu@nudt.edu.cn
ORCID: https://orcid.org/0000-0002-2315-1704

Motivation

1. On one hand, researchers focus on proposing deep neural network (DNN) robustness verification approaches. On the other hand, researchers pay attention to training robust DNNs against certain perturbations, providing guaranteed robust DNNs with proper training methods.

2. Instead of seeking a completely novel robust DNN training method, we turn our attention to alleviating the great challenge that existing interval bound propagation (IBP) family training methods encounter; i.e., IBP and CROWN-IBP methods are likely to fail under large input perturbations (drawdown risk).

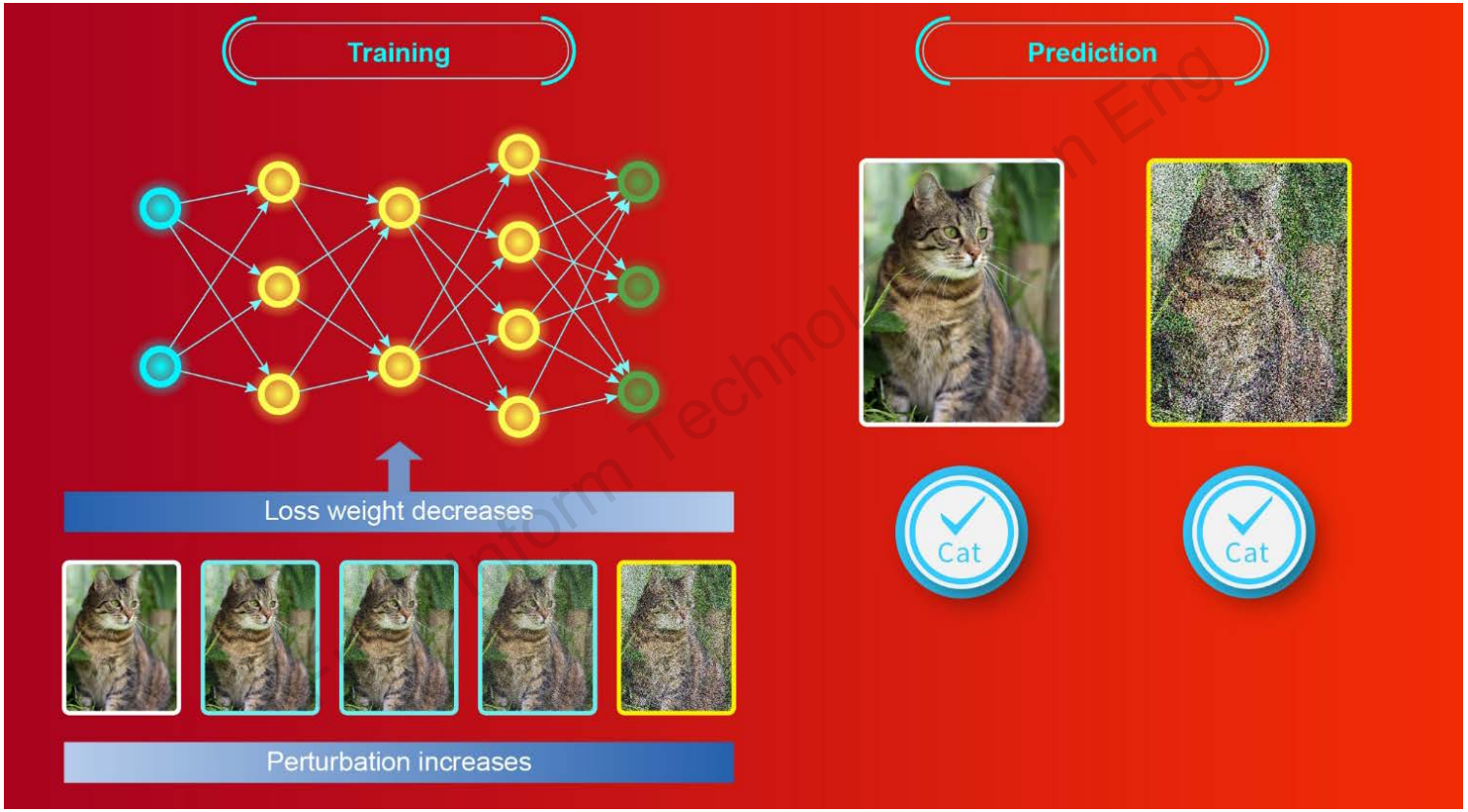
Main idea

1. To overcome the drawdown risk resulting from local robustness training (LRT), we propose the idea of global robustness training (GRT), considering some certain perturbations together, instead of only a single perturbation value, during the training processes.

2. Taking multiple perturbation values into account, we also seek a monotonically decreasing weight strategy for combining their loss values during the training phases and form the new loss function.

3. The global and monotonically decreasing robustness training strategy is compatible with existing IBP-based training methods and can be treated as a generalization of them.

Framework



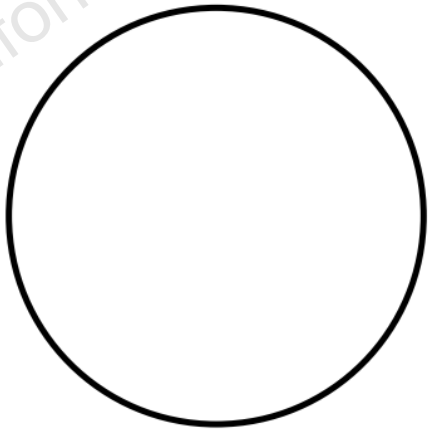
An illustrative overview of the global and monotonically decreasing robustness training strategy

Method

1. Issue: Locality leads to drawdown risk. In each training epoch, the optimization renders the parameter updating robust with respect to only a single perturbation value (i.e., the given ϵ).

$$\text{Loss}_{\text{acc}} := \mathcal{L}(\mathcal{N}(x); y; \theta) = -\sum_{(x,y)} \sum_{i=1}^{P} y_i \log(\mathcal{N}(x)_i)$$

$$\text{Loss}: = \kappa \text{Loss}_{\text{acc}} + (1-\kappa)\text{Loss}_{\text{rob}} = \kappa \mathcal{L}(\mathcal{N}(x); y; \theta) + (1-\kappa)\mathcal{L}(-\underline{m}(x, \epsilon); y; \theta)$$



Local robustness training (LRT) strategy

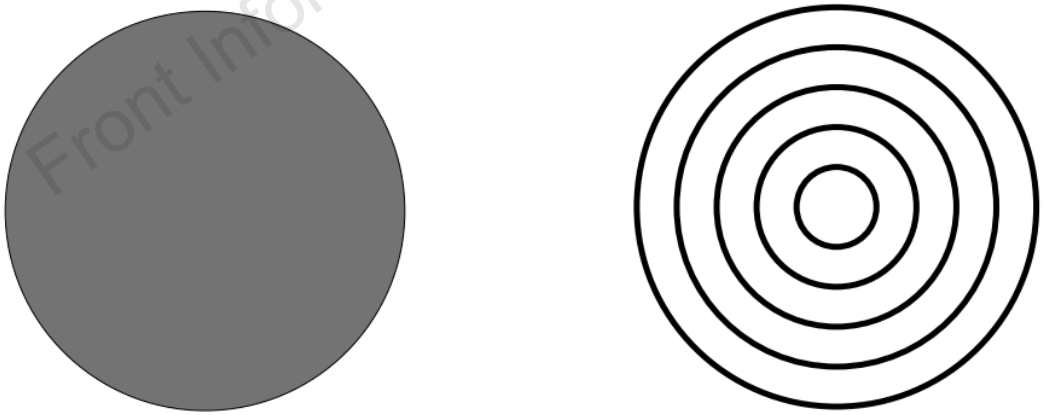
Method

2. Global robustness training strategy. Ideally, the GRT strategy refers to training with the sub-perturbation range $0 < \epsilon_{\text{train}} \leq \epsilon$ simultaneously, instead of a single perturbation ϵ.

$$\text{Loss}_{\text{rob}} = \frac{1}{\epsilon}\int_0^{\epsilon} \mathcal{L}(-\underline{m}(x, r); y; \theta)\ \text{dr}$$

It can be relaxed and approximated as

$$\text{Loss}_{\text{rob}} = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}(-\underline{m}(x, r_i); y; \theta),\ r_i = \frac{i}{N}\epsilon$$



Global robustness training (GRT) strategy
(left is the ideal case and right is the practical case)

Method

3. Monotonically decreasing robustness training strategy. To organize the different loss values, a monotonically decreasing robustness training strategy makes sense. It means that the smaller a sub-perturbation is, the more weight its loss value deserves.

$$\begin{cases} \mathrm{Loss}_{\mathrm{rob}} = \sum_{i=1}^{N} c_i \mathcal{L}(-\underline{m}(x, r_i); y; \theta) \\ r_i = \dfrac{i}{N}\epsilon, \quad c_i \propto \dfrac{1}{r_i} \end{cases}$$

Monotonically decreasing robustness training strategy

Major results

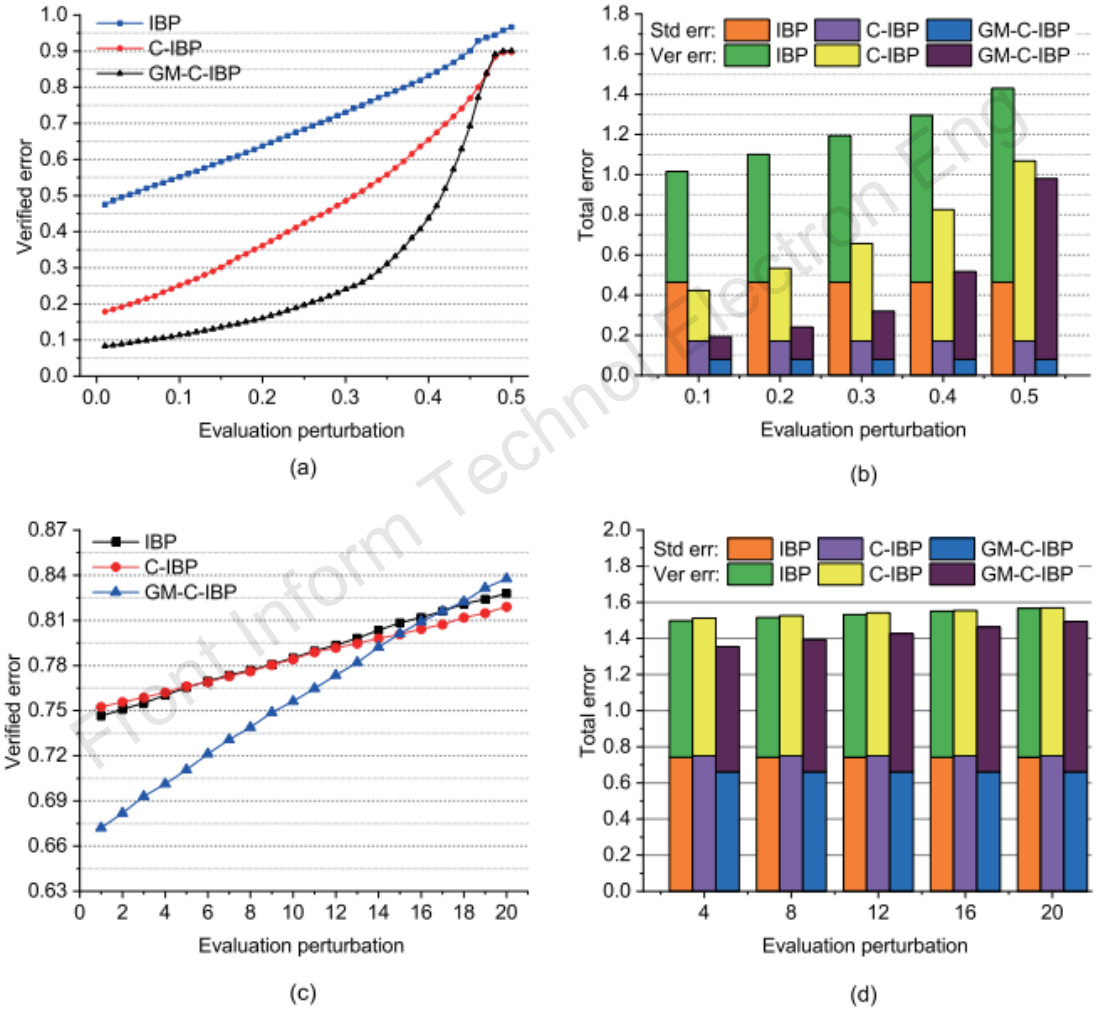


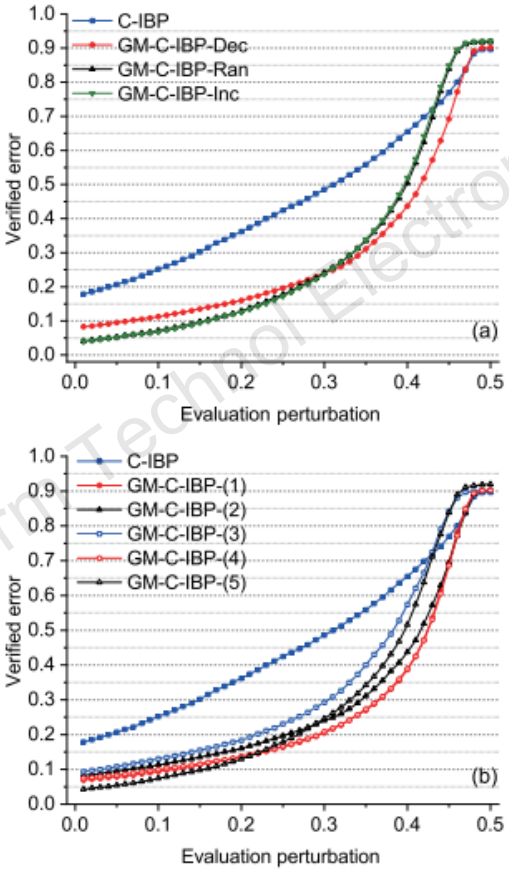
Fig. 3 Alleviating the drawdown risk: (a) verified errors on MNIST with evaluation step size 0.01; (b) total errors on MNIST with step size 0.1; (c) verified errors on CIFAR with step size 1/255; (d) total errors on CIFAR with step size 4/255

Major results

Table 3 Comparison with state-of-the-art methods against large perturbations

Dataset	Method	Standard error (%)	Verified error (%)
MNIST	IBP	46.38	62.79
$\epsilon_{\text{eval}} = 0.2$	CROWN-IBP	17.13	36.19
$\epsilon_{\text{train}} = 0.5$	GM-CROWN-IBP	**7.98**	**16.00**
MNIST	IBP	46.38	72.14
$\epsilon_{\text{eval}} = 0.3$	CROWN-IBP	17.13	48.55
$\epsilon_{\text{train}} = 0.5$	GM-CROWN-IBP	**7.98**	**24.03**
MNIST	IBP	46.38	81.92
$\epsilon_{\text{eval}} = 0.4$	CROWN-IBP	17.13	65.48
$\epsilon_{\text{train}} = 0.5$	GM-CROWN-IBP	**7.98**	**43.73**
MNIST	IBP	30.34	61.73
$\epsilon_{\text{eval}} = 0.25$	CROWN-IBP	**16.47**	48.85
$\epsilon_{\text{train}} = 0.6$	GM-CROWN-IBP	17.10	**31.25**
MNIST	IBP	30.34	88.70
$\epsilon_{\text{eval}} = 0.35$	CROWN-IBP	**16.47**	80.72
$\epsilon_{\text{train}} = 0.6$	GM-CROWN-IBP	17.10	**50.14**
CIFAR	IBP	71.77	72.39
$\epsilon_{\text{eval}} = 1/255$	CROWN-IBP	70.35	70.89
$\epsilon_{\text{train}} = 17.6/255$	GM-CROWN-IBP	**64.27**	**65.60**
CIFAR	IBP	71.77	75.93
$\epsilon_{\text{eval}} = 7/255$	CROWN-IBP	70.35	74.38
$\epsilon_{\text{train}} = 17.6/255$	GM-CROWN-IBP	**64.27**	**73.81**
CIFAR	IBP	74.23	79.82
$\epsilon_{\text{eval}} = 14/255$	CROWN-IBP	74.82	79.46
$\epsilon_{\text{train}} = 22/255$	GM-CROWN-IBP	**66.15**	**79.22**

The standard errors of the Nominal method are 1.12% and 13.80% on the MNIST and CIFAR datasets, respectively. The best results are in bold



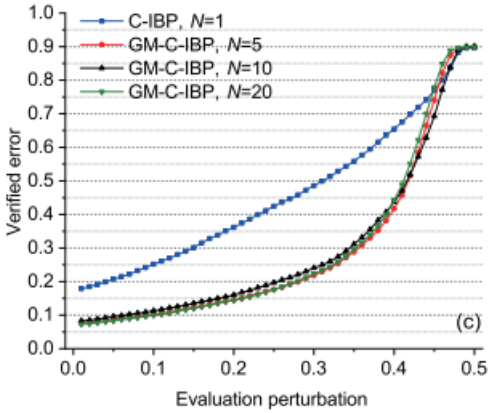


Fig. 4 Performance comparisons of GM-CROWN-IBP on different robustness weight types (a), monotonically decreasing robustness weights (b), and robustness sizes (c)

Conclusions

1. We introduce multiple sub-perturbations into the DNN training epochs to consider robustness with respect to some certain perturbations globally.

2. We organize different robustness loss values in a monotonically decreasing style to further improve the DNN robust training performance.

3. We will consider further optimization of the training duration and burdens, such as the utilization of a parallel computation mechanism and a more memory-saved estimation on the robustness margin.

Zhen LIANG received his BS degree in computer science and technology from National University of Defense Technology (NUDT), Changsha, in 2019. He is currently a PhD candidate in NUDT, Changsha. He was jointly educated in the Institute of Software, Chinese Academy of Sciences, Beijing, from 2022 to 2023. His research interests include model checking, interpretation and formal verification of artificial intelligence (AI).

Taoran WU received his BS degree in computer science and technology from Beihang University, Beijing, in 2022. He is currently a master candidate in the Institute of Software, Chinese Academy of Sciences, Beijing. His research interests involve formal verification of hybrid systems and AI.

Wanwei LIU received his PhD degree in computer science in NUDT, Changsha, in 2009. He is currently a professor in NUDT, Changsha. He is a senior member of CCF. His research interests include theoretical computer science (particularly in automata theory and temporal logic), formal methods (particularly in verification), and software engineering.

Bai XUE received his PhD degree in applied mathematics from Beihang University, Beijing, in 2014. He is currently a research professor with the Institute of Software, Chinese Academy of Sciences, Beijing, since Sept. 2021. Prior to joining the Institute of Software as an associate research professor in Nov. 2017, he worked as a research fellow with the Centre for High Performance Embedded Systems, Nanyang Technological University, from May 2014 to Sept. 2015, and as a postdoctoral reseacher with the Department für Informatik, Carl von Ossietzky Universität Oldenburg, from Nov. 2015 to Oct. 2017. His research interests involve formal verification of hybrid systems and AI.

Wenjing YANG received her PhD degree in multi-scale modeling from Manchester University, Manchester, UK. She is currently an associate research fellow in State Key Laboratory of High Performance Computing, NUDT, Changsha. Her research interests include machine learning, robotics software, and high-performance computing.

Ji WANG received his PhD degree in computer science from NUDT, Changsha, in 1995. He is currently a full professor in NUDT, Changsha, and a fellow of CCF. His research interests include software engineering and formal methods.

Zhengbin PANG received his BS, MS, and PhD degrees in computer science from NUDT, Changsha. Currently, he is a professor in NUDT, Changsha. His research interests include high-speed interconnect, heterogeneous computing, and high performance computer systems.