Certified Training: Small Boxes are All You Need

Mark Müller*  Franziska Eckert*  Marc Fischer  Martin Vechev
Standard Classification

Input $x^0$ → NN classifier → Output $y^\Lambda$

Incorrect $x$ → Correct $x$
Adversarial Examples

Szegedy et al. “Intriguing properties of neural networks” ICLR 2014
Biggio et al. “Evasion attacks against machine learning at test time” ECML PKDD 2013
Exact Propagation

Input $x^0$

$2\epsilon$

$\times$

→ FC → ReLU → \cdots → FC →

Output $y^A$

incorrect

correct

Reachable set

Certified Training: Small Boxes are All You Need – ICLR 2023 – Mark Müller – ETH Zurich
Exact Propagation

Input $x^0$

$2\epsilon$

$\rightarrow$ FC $\rightarrow$ ReLU $\rightarrow$ $\cdots$ $\rightarrow$ FC

Output $y^\Delta$

incorrect

correct

Certified Training: Small Boxes are All You Need – ICLR 2023 – Mark Müller – ETH Zurich
Standard Training
Adversarial Training (PGD)

Madry et al. "Towards Deep Learning Models Resistant to Adversarial Attacks," ICLR 2018
Adversarial Training (PGD)

Madry et al. "Towards Deep Learning Models Resistant to Adversarial Attacks." ICLR 2018
Certified Training (IBP)

Gowal, et al. "On the effectiveness of interval bound propagation for training verifiably robust models." arXiv 2018
Mirmann et al. "Differentiable abstract interpretation for provably robust neural networks." ICML 2018
SABR – This Work

Input $x^0$ → FC → ReLU → ... → FC → Output $y^\Delta$

Box relaxation

incorrect
correct
Regularisation Comparison
Worst-Case Loss Approximation Precision

Certified Training: Small Boxes are All You Need – ICLR 2023 – Mark Müller – ETH Zurich
Box Abstraction Size Growth

Growth rate \( \kappa = \frac{\mathbb{E}[\text{Output Box Size}]}{\text{Input Box Size}} \) = \[ \delta_{\text{out}} \]

\[ \delta_{\text{in}} \]
Box Abstraction Size Growth

Growth rate \( \kappa = \frac{\mathbb{E}[\text{Output Box Size}]}{\text{Input Box Size}} \)

\( \kappa \sim [10, 100] \)

Linear layers: \( \kappa \) is independent of input box scale:

\[ \delta_{\text{out}} = \delta_{\text{in}} = W \cdot \delta + b \]
Box Abstraction Size Growth

Growth rate \( \kappa = \frac{\mathbb{E}[\text{Output Box Size}]}{\text{Input Box Size}} \)

Linear layers: \( \kappa \) is independent of input box scale:
\[ \kappa \sim [10, 100] \]

ReLU layers: \( \kappa \) depends on box scale and box centre:
\[ \kappa \sim [0, 1] \]

\[ \delta_{\text{out}} = \left\{ \begin{array}{l} W + b \quad \text{if} \quad \delta_{\text{in}} > 0 \\ 0 \quad \text{otherwise} \end{array} \right. \]
Box Abstraction Size Growth

Growth rate \( \kappa = \frac{E[\text{Output Box Size}]}{\text{Input Box Size}} \) = \[ \delta_{\text{out}} \left\{ \delta_{\text{in}} \right\} \]

Linear layers: \( \kappa \) is independent of input box scale:
\[ \kappa \sim [10, 100] \]

ReLU layers: \( \kappa \) depends on box scale and box centre:
\[ \kappa \sim [0, 1] \]
Box Abstraction Size Growth – ReLUs

\[ y = \text{ReLU}(x) \]

\[ \delta_{\text{in}} \]

\[ \delta_{\text{out}} \]
Box Abstraction Size Growth – ReLUs

\[ y = \text{ReLU}(x) \]

active \implies \kappa = 1
Box Abstraction Size Growth – ReLUs

\[ y = \text{ReLU}(x) \]

inactive \( \Rightarrow \kappa = 0 \)
active \( \Rightarrow \kappa = 1 \)
Box Abstraction Size Growth – ReLUs

\[ y = \text{ReLU}(x) \]

\[ \delta_{\text{out}} \]

\[ \delta_{\text{in}} \]

inactive \( \Rightarrow \kappa = 0 \)

active \( \Rightarrow \kappa = 1 \)

crossing \( \Rightarrow \kappa = 0.5 \)

\[ \zeta = 0.5 \]

\[ \zeta = 0 \]
Box Abstraction Size Growth – ReLUs

For input box sizes $\varepsilon \to 0$

$$\kappa = \text{Portion of active ReLUs}$$

For input box sizes $\varepsilon \to \infty$

$$\kappa = 0.5$$

In-between:

$$\kappa \text{ depends on box positions}$$
Box Abstraction Size Growth – ReLUs

![Graph showing frequency distribution of pre-activation values for active and inactive neurons, with different methods indicated: IBP, SABR, PGD.](image)

Certified Training: Small Boxes are All You Need – ICLR 2023 – Mark Müller – ETH Zurich
Box Abstraction Size Growth – ReLUs

Certified Training: Small Boxes are All You Need – ICLR 2023 – Mark Müller – ETH Zurich
Box Abstraction Size Growth – ReLUs

Superlinear growth
Full Network Loss Growth

![Graph showing the growth of loss with relative box size. The x-axis represents the relative box size, ranging from 0.0 to 1.0. The y-axis represents the loss, with two scales: 10^0 and 10^1. The graph shows multiple lines representing different methods: Std, Box, IBP, and SABR.](image-url)
Impact of Verification Method

Certified Training: Small Boxes are All You Need – ICLR 2023 – Mark Müller – ETH Zurich

Accuracy [%]

Training Box Size

Certified:
- MN-BAB
- DeepPoly
- Box

more precise
Impact of Verification Method

Accuracy [%]

Training Box Size

Certified:
- MN-BAB
- DeepPoly
- Box

- Standard
- Adversarial

more precise
Impact of Verification Method

Certified: 
- Standard
- Adversarial

more precise

Accuracy [%] vs Training Box Size

Certified Training: Small Boxes are All You Need – ICLR 2023 – Mark Müller – ETH Zurich
Impact of Verification Method

Certified: Small Boxes are All You Need – ICLR 2023 – Mark Müller – ETH Zurich

- Impact of Verification Method
- Training Box Size
- Certified Training: Small Boxes are All You Need
- ICLR 2023
- Mark Müller
- ETH Zurich

more precise
Empirical Results
Empirical Results

MNIST

Cert. Acc. [%] vs Std. Acc. [%]
- $\epsilon = 0.1$
- $\epsilon = 0.3$

CIFAR-10

Cert. Acc. [%] vs Std. Acc. [%]
- $\epsilon = 2/255$
- $\epsilon = 8/255$

TinyImageNet

Cert. Acc. [%] vs Std. Acc. [%]
- $\epsilon = 1/255$

Legend:
- IBP-R
- COLT
- CROWN-IBP
- IBP
- SABR
- Ours
Conclusion
Thank You For Your Attention!

Paper & Code:

https://www.sri.inf.ethz.ch/publications/mueller2022sabr

https://github.com/eth-sri/SABR

Certified Training: Small Boxes are All You Need – ICLR 2023 – Mark Müller – ETH Zurich