Certified Training: Small Boxes are All You Need

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

Input $x^0$ \[ \times \]

NN classifier

Output $y^\Lambda$

\[ \times \checkmark \]

incorrect

\[ \times \checkmark \]

correct
Adversarial Examples

Biggio et al. "Evasion attacks against machine learning at test time" ECML PKDD 2013
Exact Propagation

Input $x^0$

\[
\begin{align*}
\text{Output } y^\Delta & \quad \text{incorrect} \\
\text{correct} & \quad \text{Reachable set}
\end{align*}
\]
Exact Propagation

Input \( x^0 \)

\[ 2\epsilon \]

\[ \times \]

\[ \rightarrow \]

\[ \text{FC} \rightarrow \text{ReLU} \rightarrow \ldots \rightarrow \text{FC} \rightarrow \]

Output \( y^\Delta \)

\[ \times \]

\[ \text{incorrect} \]

\[ \text{correct} \]
Standard Training
Adversarial Training (PGD)

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

Certified Training (IBP)

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 $\kappa = \frac{\mathbb{E} \left[ \text{Output Box Size} \right]}{\text{Input Box Size}} = \frac{\delta_{\text{out}}}{\delta_{\text{in}}}$
Box Abstraction Size Growth

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

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

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

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

Growth rate

\[ \kappa = \frac{\mathbb{E}[\text{Output Box Size}]}{\text{Input Box Size}} \]

\[ = \kappa \]

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

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

ReLU layers: \( \kappa \) depends on box scale and box centre:

\[ \kappa \sim [0, 1] \]

\[\begin{aligned}
\delta_{\text{out}} &= W + b \\
\delta_{\text{in}} &= \max(\delta_{\text{out}}, 0)
\end{aligned}\]
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:

\[ \kappa \sim [0, 1] \]

\[ \delta_{\text{out}} = W \delta_{\text{in}} + b \\
\delta_{\text{in}} = \max(\delta_{\text{in}}, 0) \]
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 \( \Rightarrow \kappa = 1 \)
Box Abstraction Size Growth – ReLUs

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

inactive \( \Rightarrow \kappa = 0 \)

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

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

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

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

inactive \( \Rightarrow \kappa = 0 \)

active \( \Rightarrow \kappa = 1 \)

crossing \( \Rightarrow \kappa = 0.5 \)
Box Abstraction Size Growth – ReLUs

For input box sizes $\varepsilon \rightarrow 0$

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

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

$$\kappa = 0.5$$

In-between:

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

![Graph showing frequency distribution](image)

- **inactive**
- **active**

Legend:
- IBP
- SABR
- PGD

Pre-Activation Value $x$
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 loss growth with relative box size]
Impact of Verification Method

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

Certified:
- Standard
- Adversarial

Accuracy [%]

Training Box Size

more precise
Impact of Verification Method

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

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

more precise
Impact of Verification Method

Accuracy [%]

Training Box Size

Certified:
- MN-BAB
- DeepPoly
- Box

Standard
Adversarial

more precise
Empirical Results

Cert. Acc. [%]

Std. Acc. [%]

better
Empirical Results

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

**CIFAR-10**
- Cert. Acc. [%]
- Std. Acc. [%]
- $\epsilon = 2/255$
- $\epsilon = 8/255$

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

- 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