Certifiable Neural Networks

safeai.ethz.ch

Matthew Mirman

February 5, 2021

SafeAI - safeai.ethz.ch

Collaborators

Martin Vechev

Timon Gehr

Maximilian Baader

Petar Tsankov

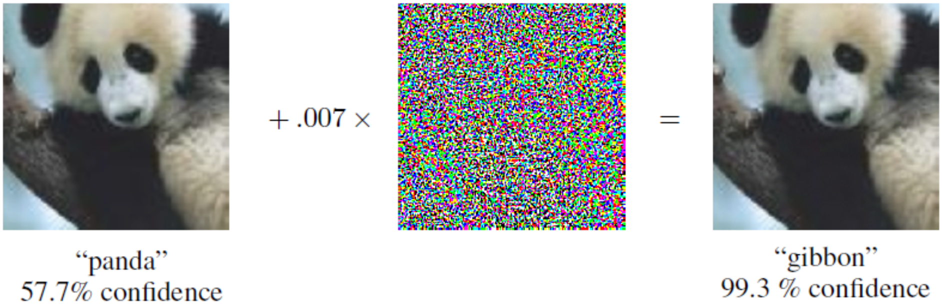
Gagandeep Singh

Dana Drachsler

Publications (discussed here):

- ICML'18 - Differentiable Abstract Interpretation for Provably Robust Neural Networks (DiffAI)
- ICLR'20 - Universal Approximation with Certified Networks
- NeurIPS'18 - Fast and Effective Robustness Certification (ERAN)
- S&P'18 - AI2: Safety and Robustness Certification of Neural Networks with Abstract Interpretation

Motivation - Adversarial Attacks



Example of FGSM attack produced by Goodfellow et al. (2014)

Overview of Neural Network Safety

Identifying vulnerability

Adversarial attacks: Goodfellow et al. (2014)

Reducing vulnerability

Adversarial training: Madry et al. (2018)

Overview of Neural Network Safety

Identifying vulnerability

Adversarial attacks: Goodfellow et al. (2014)

Reducing vulnerability

Adversarial training: Madry et al. (2018)

Certifying invulnerability

Robustness certification: AI2, ERAN

Increasing certifiable invulnerability

Certifiable training: DiffAI

Overview of Neural Network Safety

Identifying vulnerability

Adversarial attacks: Goodfellow et al. (2014)

Reducing vulnerability

Adversarial training: Madry et al. (2018)

Certifying invulnerability

Robustness certification: AI2, ERAN

Increasing certifiable invulnerability

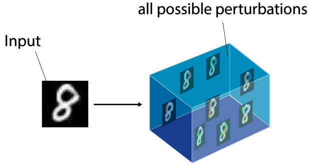
Certifiable training: DiffAI

AI2 - First scalable verifier for neural networks.

DiffAI - First scalable framework for certifiable training.

Preliminary Property - L_∞ Adversarial Ball

Many developed attacks: Goodfellow et al. (2014); Madry et al. (2018); Evtimov et al. (2017); Athalye & Sutskever (2017); Papernot et al. (2017); Xiao et al. (2018); Carlini & Wagner (2017); Yuan et al. (2017); Tramèr et al. (2017)



$$\mathrm{Ball}_\epsilon(input) = \{attack \mid \|input - attack\|_\infty \leqslant \epsilon\}$$

A net is *ϵ-robust* at x if it classifies every example in $\mathrm{Ball}_\epsilon(x)$ the same and correctly

Certification

Robustness Certification

Verification: Prove that a network is ϵ-robust at a point

Abstract Interpretation: certify by over-approximation [Cousot & Cousot (1977)]

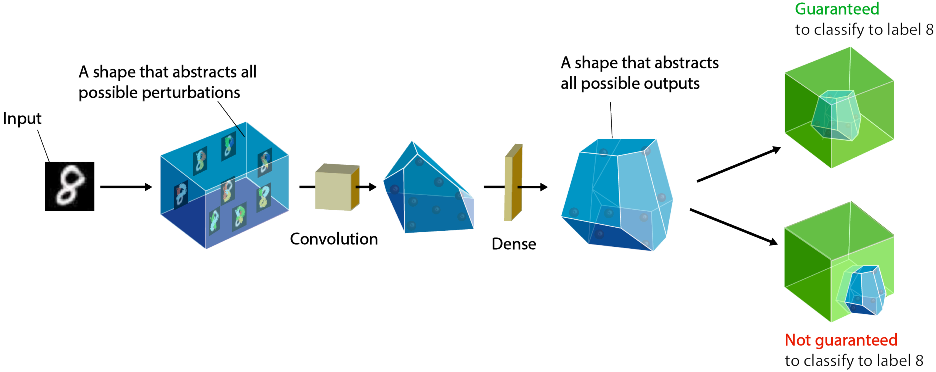
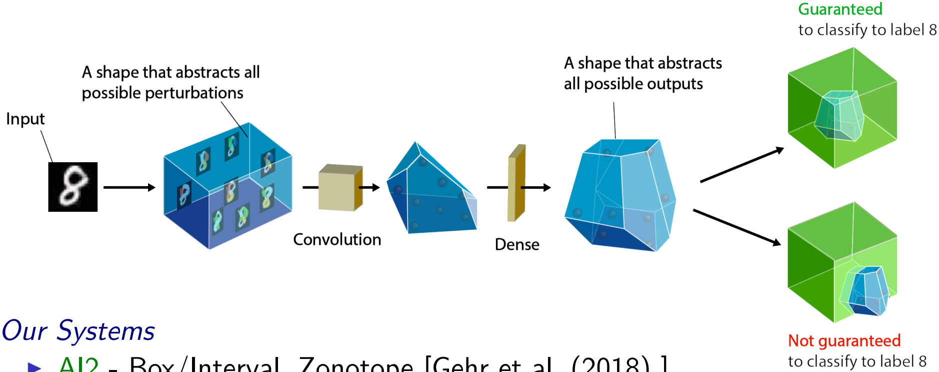


Image Credit: Petar Tsankov

Robustness Certification

Verification: Prove that a network is ϵ-robust at a point

Abstract Interpretation: certify by over-approximation [Cousot & Cousot (1977)]



Our Systems

- AI2 - Box/Interval, Zonotope [Gehr et al. (2018)]
- DiffAI - HybridZono, zBox, zDiag, zSwitch, zSmooth [Mirman et al. (2018)]
- ERAN - DeepZono [Singh et al. (2018)]

Image Credit: Petar Tsankov

Abstract Interpretation

Cousot & Cousot (1977)

Abstract Interpretation is heavily used in industrial large-scale program analysis to compute over-approximation of program behaviors [1]

Provide

- domain $\mathcal{D}$ of abstract objects d
- concretization function $\gamma : \mathcal{D} \to \mathcal{P}(\mathbb{R}^n)$
- concrete function $f : \mathbb{R}^n \to \mathbb{R}^n$

Develop a sound transformer $f^{\#} : \mathcal{D} \to \mathcal{D}$

[1] For example by Astrée: Blanchet et al. (2003)

Abstract Interpretation

Cousot & Cousot (1977)

Abstract Interpretation is heavily used in industrial large-scale program analysis to compute over-approximation of program behaviors [1]

Provide

- domain $\mathcal{D}$ of abstract objects d
- concretization function $\gamma : \mathcal{D} \to \mathcal{P}(\mathbb{R}^n)$
- concrete function $f : \mathbb{R}^n \to \mathbb{R}^n$

Develop a sound transformer $f^{\#} : \mathcal{D} \to \mathcal{D}$

- ReLU $: \mathbb{R}^n \to \mathbb{R}^n$ becomes
  $\text{ReLU}^{\#} : \mathcal{D} \to \mathcal{D}$

[1] For example by Astrée: Blanchet et al. (2003)

Abstract Interpretation

Cousot & Cousot (1977)

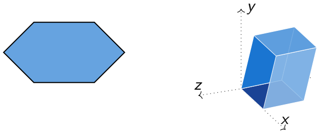
Abstract Interpretation is heavily used in industrial large-scale program analysis to compute over-approximation of program behaviors [1]

Provide

- domain $\mathcal{D}$ of abstract objects d
- concretization function $\gamma : \mathcal{D} \to \mathcal{P}(\mathbb{R}^n)$
- concrete function $f : \mathbb{R}^n \to \mathbb{R}^n$

Develop a sound transformer $f^{\#} : \mathcal{D} \to \mathcal{D}$

- ReLU $: \mathbb{R}^n \to \mathbb{R}^n$ becomes ReLU$^{\#} : \mathcal{D} \to \mathcal{D}$



Zonotope Domain

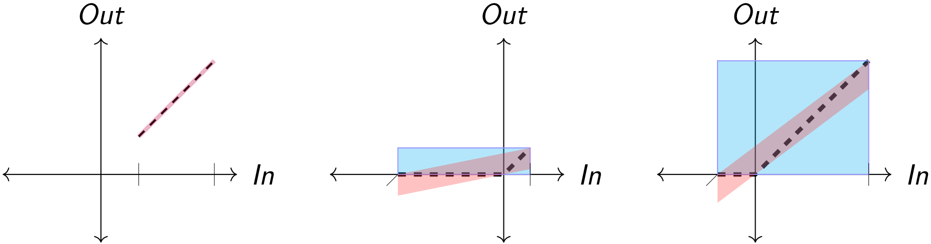
- Affine transform of k-cube onto p dims
- Ball$_\epsilon$: perfect
- $(\cdot M)^{\#}$: perfect
- ReLU$^{\#}$: zBox, zDiag, DeepZono

[1] For example by Astrée: Blanchet et al. (2003)

Zonotope Domain

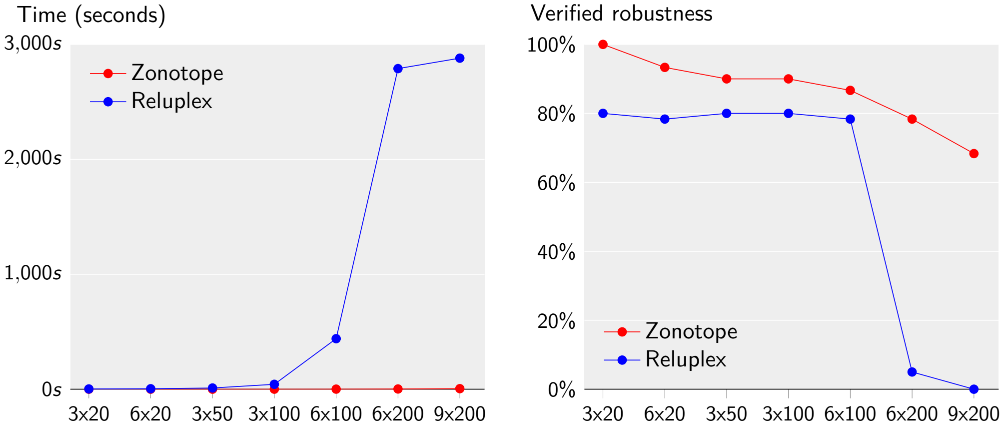
Example ReLU Transformers for Zonotope

Examples of zBox (blue) and DeepZono (red) for approximating $Out = ReLU(In)$ (dashed line).



- zBox: Treat as Box when surrounding zero
- DeepZono: Minimize area in In/Out plane.

AI2 Certification Results



Comparison to Reluplex, Katz et al. (2017), on small feed-forward networks for MNIST.

Training

Reducing Vulnerability

Certification Caveat

- Neural networks aren't robust by default.
- Why try to certify non-robust networks?

Adversarial Training

Defense: Train a network so that *most* inputs are *mostly* robust.

- Madry et al. (2018); Tramèr et al. (2017); Cisse et al. (2017); Yuan et al. (2017)

Certifiable Training

- Experimentally robust nets not necessarily *certifiably* robust
- Intuition: not all correct programs are easily provable

Certifiable Training

Train a Network to be *Certifiably* Robust[2]

Given:

- Net_θ with weights θ
- Training inputs and labels

Find:

- θ that maximizes number of inputs we can *certify* are ϵ-robust

[2] Also addressed by: Raghunathan et al. (2018); Kolter & Wong (2017); Dvijotham et al. (2018)

Certifiable Training

Train a Network to be *Certifiably* Robust[2]

Given:

- Net_θ with weights θ
- Training inputs and labels

Find:

- θ that maximizes number of inputs we can *certify* are ϵ-robust

Method

- Design a loss function based on certification goal
- Differentiate through certifier
- Perform SGD

[2] Also addressed by: Raghunathan et al. (2018); Kolter & Wong (2017); Dvijotham et al. (2018)

Scalability
CIFAR10

			Time to Train 1 Epoch	
Model	#Neurons	#Weights	Baseline	DiffAI
SkipNet-18[3]	∼558k	∼16mill	152s	260s

- Can use a less precise domain for training than for certification
- Can test and train with larger nets than prior work

[3]like that described by He et al. (2016) but without pooling or dropout.

Robustness Provability

CIFAR10 with $\epsilon = 0.012$ [4]

Training Method	%Certified DeepZono
Baseline	0
Adversarial	7
DiffAI	64

- Significantly increases provability with scalable verifiers.
- For small ϵ we lose little accuracy.

[4] Numbers from Singh et al. (2018) on 100 test images

The Gap

The state of the art still far from goal

- Balunovic & Vechev (2019) gets 60.5% certified robustness and 78.4% accuracy on CIFAR10 with $\epsilon = \frac{2}{255}$
- Standard training $>$ 95% accuracy.

[5] Baader M, Mirman M, Vechev M. Universal Approximation with Certified Networks. In ICLR 2020

The Gap

The state of the art still far from goal

- Balunovic & Vechev (2019) gets 60.5% certified robustness and 78.4% accuracy on CIFAR10 with $\epsilon = \frac{2}{255}$
- Standard training $> 95\%$ accuracy.

- Universal approximation implies robust networks exist.
- Network verification is NP-complete in general.
- Do robust *and* convexly certifiable networks exist?

[5] Baader M, Mirman M, Vechev M. Universal Approximation with Certified Networks. In ICLR 2020

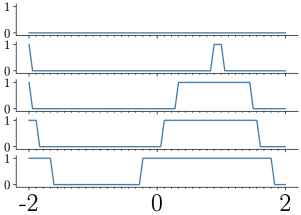
The Gap

The state of the art still far from goal

- Balunovic & Vechev (2019) gets 60.5% certified robustness and 78.4% accuracy on CIFAR10 with $\epsilon = \frac{2}{255}$
- Standard training $>$ 95% accuracy.

- Universal approximation implies robust networks exist.
- Network verification is NP-complete in general.
- Do robust *and* convexly certifiable networks exist?

They provably exist![5]



[5] Baader M, Mirman M, Vechev M. Universal Approximation with Certified Networks. In ICLR 2020

Universal Approximation with Certified Networks

Universal Interval-Certified Approximation

Let $\Gamma \subset \mathbb{R}^m$ be a compact set and let $f : \Gamma \to \mathbb{R}$ be a continuous function. For all $\delta > 0$, there exists a ReLU network n such that for all boxes $[a, b]$ in Γ defined by points $a, b \in \Gamma$ where $a_k \leq b_k$ for all k, the propagation of the box $[a, b]$ using interval analysis through the network n, denoted $n^\sharp([a, b])$, approximates the set $[l, u] = [\min f([a, b]), \max f([a, b])] \subseteq \mathbb{R}$ up to δ,

$$[l + \delta, u - \delta] \subseteq n^\sharp([a, b]) \subseteq [l - \delta, u + \delta]$$

tldr: univeral approximation can be lifted to nets that are certifiably robust with *Box*.

Universal Approximation with Certified Networks

Universal Interval-Certified Approximation

Let $\Gamma \subset \mathbb{R}^m$ be a compact set and let $f : \Gamma \to \mathbb{R}$ be a continuous function. For all $\delta > 0$, there exists a ReLU network n such that for all boxes $[a, b]$ in Γ defined by points $a, b \in \Gamma$ where $a_k \leq b_k$ for all k, the propagation of the box $[a, b]$ using interval analysis through the network n, denoted $n^\sharp([a, b])$, approximates the set $[l, u] = [\min f([a, b]), \max f([a, b])] \subseteq \mathbb{R}$ up to δ,

$$[l + \delta, u - \delta] \subseteq n^\sharp([a, b]) \subseteq [l - \delta, u + \delta]$$

tldr: univeral approximation can be lifted to nets that are certifiably robust with *Box*.

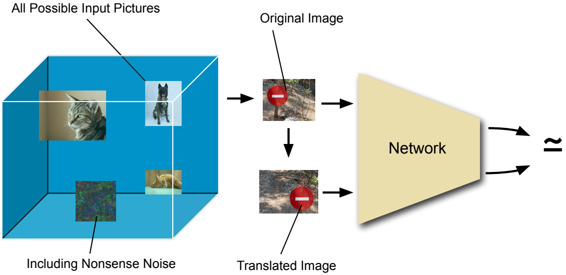
Future Work

Describe lower & upper bounds on size/depth/width with certifiable networks

Beyond Local Robustness

Network Invariants (Future Work)

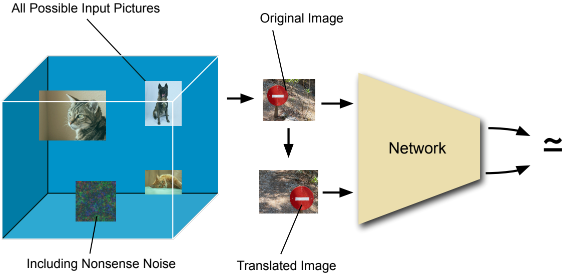
- Verify properties across every possible input.
- Invariance to translations, rotations, arbitrary perturbations.

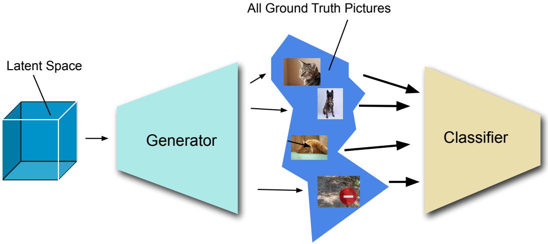


Beyond Local Robustness

Network Invariants (Future Work)

- Verify properties across every possible input.
- Invariance to translations, rotations, arbitrary perturbations.





Generative Specifications (Ongoing)

Specify input with generative models [Mirman et al. (2020)]

Generative Specifications

- Generative specifications are necessarily non-convex.
- Feasible Restriction: Interpolative Specifications

Our Work

certify images from interpolations in
the generative model's latent space

Convex Relaxation Analysis

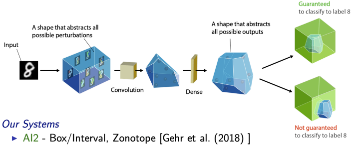
analyzes non-realistic images
from pixel-wise interpolation

Conclusion

First scalable certification framework

Robustness Certification

Verification: Prove that a network is ϵ-robust at a point
Abstract Interpretation: certify by over-approximation [Cousot & Cousot (1977)]



Our Systems

- AI2 - Box/Interval, Zonotope [Gehr et al. (2018)]
- DiffAI - HybridZono, zBox, zDiag, zSwitch, zSmooth [Mirman et al. (2018)]
- ERAN - DeepZono [Singh et al. (2018)]

Image Credit: Petar Tsankov

Existence of interval provable nets

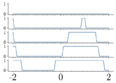
The Gap

The state of the art still far from goal

- Balunovic & Vechev (2019) gets 60.5% certified robustness and 78.4% accuracy on CIFAR10 with $\epsilon = \frac{2}{255}$
- Standard training $>$ 95% accuracy.

- Universal approximation implies robust networks exist.
- Network verification is NP-complete in general.
- Do robust *and* convexly certifiable networks exist?

They provably exist![5]



[5]Baader M, Mirman M, Vechev M. Universal Approximation with Certified Networks. In ICLR 2020

First scalable certification training framework

Scalability

CIFAR10

Model	#Neurons	#Weights	Time to Train 1 Epoch	
			Baseline	DiffAI
SkipNet-18[3]	∼558k	∼16mill	152s	260s

- Can use a less precise domain for training than for certification
- Can test and train with larger nets than prior work

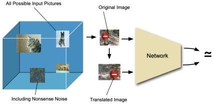
[3]like that described by He et al. (2016) but without pooling or dropout.

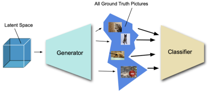
Going beyond local robustness

Beyond Local Robustness

Network Invariants (Future Work)

- Verify properties across every possible input.
- Invariance to translations, rotations, arbitrary perturbations.





Generative Specifications (Ongoing)

Specify input with generative models [Mirman et al. (2020)]

Bibliography I

Athalye, A. and Sutskever, I. Synthesizing robust adversarial examples. *arXiv preprint arXiv:1707.07397*, 2017.

Balunovic, M. and Vechev, M. Adversarial training and provable defenses: Bridging the gap. In *International Conference on Learning Representations*, 2019.

Blanchet, B., Cousot, P., Cousot, R., Feret, J., Mauborgne, L., Miné, A., Monniaux, D., and Rival, X. A static analyzer for large safety-critical software. In *Programming Language Design and Implementation (PLDI)*, 2003.

Carlini, N. and Wagner, D. A. Adversarial examples are not easily detected: Bypassing ten detection methods. *CoRR*, abs/1705.07263, 2017. URL `http://arxiv.org/abs/1705.07263`.

Cisse, M., Bojanowski, P., Grave, E., Dauphin, Y., and Usunier, N. Parseval networks: Improving robustness to adversarial examples. In *International Conference on Machine Learning*, pp. 854–863, 2017.

Bibliography II

Cousot, P. and Cousot, R. Abstract interpretation: a unified lattice model for static analysis of programs by construction or approximation of fixpoints. In *Symposium on Principles of Programming Languages (POPL)*, 1977.

Dvijotham, K., Gowal, S., Stanforth, R., Arandjelovic, R., O'Donoghue, B., Uesato, J., and Kohli, P. Training verified learners with learned verifiers. *arXiv preprint arXiv:1805.10265*, 2018.

Evtimov, I., Eykholt, K., Fernandes, E., Kohno, T., Li, B., Prakash, A., Rahmati, A., and Song, D. Robust physical-world attacks on deep learning models. *arXiv preprint arXiv:1707.08945*, 2017.

Gehr, T., Mirman, M., Tsankov, P., Drachsler Cohen, D., Vechev, M., and Chaudhuri, S. Ai2: Safety and robustness certification of neural networks with abstract interpretation. In *Symposium on Security and Privacy (SP)*, 2018.

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

Bibliography III

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

Katz, G., Barrett, C., Dill, D. L., Julian, K., and Kochenderfer, M. J. Reluplex: An efficient smt solver for verifying deep neural networks. In *International Conference on Computer Aided Verification*, 2017.

Kolter, J. Z. and Wong, E. Provable defenses against adversarial examples via the convex outer adversarial polytope. *arXiv preprint arXiv:1711.00851*, 2017.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. 2018.

Mirman, M., Gehr, T., and Vechev, M. Differentiable abstract interpretation for provably robust neural networks. In *International Conference on Machine Learning (ICML)*, 2018.

Mirman, M., Gehr, T., and Vechev, M. Robustness certification of generative models. *arXiv preprint arXiv:2004.14756*, 2020.

Bibliography IV

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., and Swami, A. Practical black-box attacks against machine learning. In *Asia Conference on Computer and Communications Security*. ACM, 2017.

Raghunathan, A., Steinhardt, J., and Liang, P. Certified defenses against adversarial examples. *arXiv preprint arXiv:1801.09344*, 2018.

Singh, G., Gehr, T., Mirman, M., Püschel, M., and Vechev, M. Fast and effective robustness certification. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 10825–10836, 2018.

Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., and McDaniel, P. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

Xiao, C., Li, B., Zhu, J.-Y., He, W., Liu, M., and Song, D. Generating adversarial examples with adversarial networks. *arXiv preprint arXiv:1801.02610*, 2018.

Yuan, X., He, P., Zhu, Q., Bhat, R. R., and Li, X. Adversarial examples: Attacks and defenses for deep learning. *arXiv preprint arXiv:1712.07107*, 2017.