


# First three years of the international verification of neural networks competition (VNN-COMP)

**Journal Article****Author(s):**

Brix, Christopher; [Müller, Mark Niklas](#) ; Bak, Stanley; Johnson, Taylor T.; Liu, Changliu

**Publication date:**

2023-06

**Permanent link:**

<https://doi.org/10.3929/ethz-b-000617162>

**Rights / license:**

[Creative Commons Attribution 4.0 International](#)

**Originally published in:**

International Journal on Software Tools for Technology Transfer 25(3), <https://doi.org/10.1007/s10009-023-00703-4>



# First three years of the international verification of neural networks competition (VNN-COMP)

Christopher Brix<sup>1</sup> · Mark Niklas Müller<sup>2</sup> · Stanley Bak<sup>3</sup> · Taylor T. Johnson<sup>4</sup> · Changliu Liu<sup>5</sup>

Accepted: 14 January 2023 / Published online: 30 May 2023  
© The Author(s) 2023

## Abstract

This paper presents a summary and meta-analysis of the first three iterations of the annual International Verification of Neural Networks Competition (VNN-COMP), held in 2020, 2021, and 2022. In the VNN-COMP, participants submit software tools that analyze whether given neural networks satisfy specifications describing their input-output behavior. These neural networks and specifications cover a variety of problem classes and tasks, corresponding to safety and robustness properties in image classification, neural control, reinforcement learning, and autonomous systems. We summarize the key processes, rules, and results, present trends observed over the last three years, and provide an outlook into possible future developments.

**Keywords** Certified robustness · Adversarial robustness · Formal verification · Formal methods · Neural networks · Machine learning · Deep learning

## 1 Introduction

Neural networks are increasingly used in safety-critical applications [7, 24]. However, it has become apparent that they are highly susceptible to adversarial examples [52], i.e., minor and possibly imperceptible input perturbations can cause the output to change significantly. As such perturbations can occur in the real world either at random or due to malicious actors, it is of utmost importance to analyze the robustness of deep learning based systems in a mathematically rigorous manner before applying them in safety-critical domains. To

this end, a wide range of methods and corresponding software tools have been developed [13, 17, 23, 25]. However, with tools becoming ever more numerous and specialized, it became increasingly difficult for practitioners to decide which tool to use.

In 2020, the inaugural VNN-COMP was organized to tackle this problem and allow researchers to compare their neural network verifiers on a wide set of benchmarks. Initially conceived as a friendly competition with little standardization, it was increasingly standardized and automated to ensure a fair comparison on cost-equivalent hardware using standardized formats for both properties and networks.

In this work, we outline this development, summarize key rules and results, describe the high-level trends observed over the last three years, and provide an outlook on possible future developments.

## 2 Neural network verification

We consider the neural network verification problem defined as follows: Given an input specification  $\phi \subseteq \mathbb{R}^{d_{in}}$ , also called pre-condition, an output specification  $\psi \subseteq \mathbb{R}^{d_{out}}$ , also called post-condition, and a neural network  $N : \mathbb{R}^{d_{in}} \mapsto \mathbb{R}^{d_{out}}$ , we aim to prove that the pre-condition implies the post-condition, i.e.,

$$\forall x : x \models \phi \Rightarrow N(x) \models \psi, \quad (1)$$

✉ C. Brix  
[brix@cs.rwth-aachen.de](mailto:brix@cs.rwth-aachen.de)  
M.N. Müller  
[mark.mueller@inf.ethz.ch](mailto:mark.mueller@inf.ethz.ch)  
S. Bak  
[stanley.bak@stonybrook.edu](mailto:stanley.bak@stonybrook.edu)  
T.T. Johnson  
[taylor.johnson@vanderbilt.edu](mailto:taylor.johnson@vanderbilt.edu)  
C. Liu  
[cliu6@andrew.cmu.edu](mailto:cliu6@andrew.cmu.edu)

<sup>1</sup> RWTH Aachen University, Aachen, Germany  
<sup>2</sup> ETH Zurich, Zurich, Switzerland  
<sup>3</sup> Stony Brook University, Stony Brook, NY, USA  
<sup>4</sup> Vanderbilt University, Nashville, TN, USA  
<sup>5</sup> Carnegie Mellon University, Pittsburgh, PA, USA

or provide a counterexample.

Inspired by the notation common in the SAT-solver community, we encode this problem by specifying a constraint set describing an adversarial example, i.e.,

$$\exists x : x \models \phi \wedge N(x) \models \neg\psi. \quad (2)$$

Therefore, we call instances where Equation (2) is satisfiable and thus the property encoded by Equation (1) does *not* hold SAT, and instances where Equation (2) is unsatisfiable and the property encoded by Equation (1) has been shown to hold UNSAT. Note that while it is possible to show SAT by directly searching for counter-examples using adversarial attacks [18, 34], these approaches are not complete, i.e., if they are not successful in finding a counter-example this does *not* imply that a property holds.

**Example problems** One particularly popular property is the robustness to adversarial  $\ell_\infty$ -norm bounded perturbations in image classification. There, the network  $N$  computes a numerical score  $y \in \mathbb{R}^{d_{\text{out}}}$  corresponding to its confidence that the input belongs to each of the  $d_{\text{out}}$  classes for each input  $x \in \mathbb{R}^{d_{\text{in}}}$ . The final classification  $c$  is then computed as  $c = \arg \max_i N(x)_i$ . In this setting, an adversary may want to perturb the input such that the classification changes. Therefore, the verification intends to prove that

$$\begin{aligned} \arg \max_i N(x')_i &= t, \\ \forall x' \in \{x' \in \mathbb{R}^{d_{\text{in}}} \mid \|x - x'\|_\infty \leq \epsilon\}, \end{aligned}$$

where  $t$  is the target class,  $x$  is the original image, and  $\epsilon$  is the maximal permissible perturbation magnitude. There, the pre-condition  $\phi$  describes the inputs an attacker can choose from ( $\phi = \{x' \in \mathbb{R}^{d_{\text{in}}} \mid \|x - x'\|_\infty \leq \epsilon\}$ ), i.e., an  $\ell_\infty$ -ball of radius  $\epsilon$ , and the post-condition  $\psi$  describes the output space corresponding to a classification to the target class  $t$  ( $\psi = \{y \in \mathbb{R}^{d_{\text{out}}} \mid y_t > y_i, \forall i \neq t\}$ ).

When neural networks are used as controllers, more complex properties can be relevant. For example, in the ACAS Xu setting [24], a neural controller gives action recommendations based on the relative position and heading of the controlled and intruder aircraft. There, we want to, e.g., ensure that for inputs  $\mathcal{D}$  corresponding to the intruder aircraft being straight ahead and heading our way, neither of the evasive maneuvers “strong left” (SL) or “strong right” (SR) is considered the worst option. More formally, we want to verify that

$$\arg \min_i N(x')_i \notin \{\text{SL}, \text{SR}\}, \forall x' \in \mathcal{D}.$$

Here, we obtain a more complex, non-convex post-condition

$$\begin{aligned} \psi &= \mathbb{R}^{d_{\text{out}}} \setminus \\ &\quad (\{y \in \mathbb{R}^{d_{\text{out}}} \mid y_{\text{SL}} < y_i, \forall i \notin \{\text{SL}, \text{SR}\}\} \\ &\quad \cup \{y \in \mathbb{R}^{d_{\text{out}}} \mid y_{\text{SR}} < y_i, \forall i \notin \{\text{SL}, \text{SR}\}\}). \end{aligned}$$

### 3 Competition goals

VNN-COMP is organized to further the following goals.

**Define standards** To enable practitioners to easily use and evaluate a range of different verification approaches and tools without substantial overhead, it is essential that all tools can process both networks and specifications in a standardized file format. To this end, the second iteration of the VNN-COMP established such a standard. Problem specifications (pre- and post-condition) are defined using the VNN-LIB [53] format and neural networks are defined using the ONNX [3] standard. In 2022, additionally, a standardized format for counterexamples was introduced.

**Facilitate verification tool comparison** Every year, dozens of papers are published on neural network verification, many proposing not only new methods but also new benchmarks. With authors potentially investing more time into tuning their method to the chosen benchmarks, a fair comparison between all these methods is difficult. VNN-COMP facilitates such a comparison between a large number of tools on a diverse set of benchmarks, using cost-equivalent hardware, and test instances not available to participants. Letting participants and industry practitioners propose a wide range of interesting benchmarks, yields not only a ranking on the problems typically used in the field, but also highlights which tools are particularly suitable for more specialized problems. Further, by ensuring a standardized installation and evaluation process is in place, the comparison to a large number of state-of-the-art tools for any publication is enabled.

**Shape future work directions** The visibility VNN-COMP lends to the problems underlying the considered benchmarks has the potential to raise their profile in the community. As benchmarks are developed jointly by industry and academia, this constitutes a great opportunity to shape future research to be as impactful as possible. Over the last years, benchmarks have featured ever-increasing network sizes (see Table 5), promoting scalability, more complex networks (including, e.g., residual [20] and max-pooling layers [69]), promoting generalizability, and more complex specifications, enabling more interesting properties to be analyzed.

**Bring researchers together** Both the rule and benchmark discussion phase during the lead-up to the competition, as well as the in-person presentation of results at the Workshop on Formal Methods for ML-Enabled Autonomous Systems (FoMLAS)<sup>1</sup> provide participants with a great opportunity to meet fellow researchers and discuss the future of the field. Further, the tool and benchmark descriptions participants provide for the yearly report [2, 5, 36] serve as an excellent summary of state-of-the-art methods, allowing people entering the field to get a quick overview.

## 4 Overview of three years of VNN-COMP

In this section, we provide a high-level description of how the VNN-COMP evolved from 2020 to 2022, listing all participants and the final rankings in Table 4. Generally, performance is measured on a set of equally weighted *benchmarks*, each consisting of a set of related *instances*. Each instance consists of a trained neural network, a timeout, and input and output constraints. Below, we group benchmarks into *categories* to enable a quicker comparison between years.

### 4.1 VNN-COMP 2020

The inaugural VNN-COMP<sup>2</sup> [2] was held in 2020 as a “friendly competition” with no winner. Its main goal was to provide a stepping stone for future iterations by starting the process of defining common problem settings and identifying possible avenues for standardization.

#### 4.1.1 Benchmarks

Three benchmark categories were considered with only one of the eight teams participating in all of them:

- Fully connected networks with ReLU activations – two benchmarks, based on ACAS Xu and MNIST.
- Fully connected networks with sigmoid and tanh activation functions – one benchmark, based on MNIST.
- Convolutional networks – two benchmarks, based on MNIST and CIFAR10.

#### 4.1.2 Evaluation

Teams evaluated their tools using their own hardware. While this simplified the evaluation process, it made the reported results incomparable, due to the significant hardware differences. The teams reported that they used between 4 and 40 CPUs and between 16 and 756 GB of RAM.

### 4.2 VNN-COMP 2021

Based upon the insights gained in 2020, the second iteration of VNN-COMP<sup>3</sup> was organized with a stronger focus on comparability between the participating tools [5].

#### 4.2.1 Benchmarks

Teams were permitted to propose one benchmark with a total timeout of at most six hours, split over its constituting instances. Networks were defined in the ONNX format [3] and problem specifications were given in the VNN-LIB format [53]. To prevent excessive tuning to specific benchmark instances, benchmark proposers were encouraged to provide a script enabling the generation of new random instances for the final tool evaluation. However, teams were allowed to tune their tools for each benchmark, using the initial set of benchmark instances.

In 2021, the benchmarks could be split into the following categories, with multiple teams participating in all of them:

- Fully connected networks with ReLU activations – two benchmarks, based on ACAS Xu and MNIST.
- Fully connected networks with sigmoid activations – one benchmark, based on MNIST.
- Convolutional networks – three benchmarks, based on CIFAR10.
- Networks with max-pooling layers – one benchmark, based on MNIST.
- Residual networks – one benchmark, based on CIFAR10.
- Large networks with sparse matrices – one benchmark, based on database indexing.

#### 4.2.2 Evaluation

To allow for comparability of results, all tools were evaluated on equal-cost hardware using Amazon Web Services (AWS). Each team could decide whether they wanted their tool to be evaluated on a CPU-focused r5.12xlarge or a GPU-focused p3.2xlarge instance (see Table 1 for more details). Further, instead of providing results and runtimes themselves, teams had to prepare scripts automating the installation and execution of their tools. After the submission deadline, the organizers installed and evaluated each tool using the provided scripts. In many cases, this process required some debugging in a back and forth between the organizers and teams.

**Scoring** For every benchmark, 10 points were awarded for correctly showing the instance to be SAT/UNSAT, with a 100 point penalty for incorrect results (see Table 2). A simple adversarial attack was used to identify “easy” SAT instances,

<sup>1</sup> <https://fomlas2022.wixsite.com/fomlas2022>.

<sup>2</sup> <https://sites.google.com/view/vnn20/vnncomp>.

<sup>3</sup> <https://sites.google.com/view/vnn2021>.

**Table 1** Available AWS instances

	2021	2022	vCPUs	RAM [GB]	GPU
r5.12xlarge	✓	✗	48	384	✗
p3.2xlarge	✓	✓	8	61	V100 GPU with 16 GB memory
m5.16xlarge	✗	✓	64	256	✗
g5.8xlarge	✗	✓	32	128	A10G GPU with 24 GB memory
t2.large	✗	✓	2	8	✗

**Table 2** Points per instance in 2021. SAT instances were split into simple and complex, based on whether a simple adversarial attack was successful

Ground truth	Returned result		
	SAT	UNSAT	Other
SAT, simple	+1	−100	0
SAT, complex	+10	−100	0
UNSAT	−100	+10	0

on which the available points were reduced from 10 to 1. If tools reported contradicting results on an instance, the ground truth was decided by a majority vote. Bonus points were awarded to the fastest two tools on every instance (two points for the fastest and one point for the second fastest). Runtimes differing by less than 0.2 seconds or below one second were considered equal, so multiple teams could receive the two point bonus. To correct the notable differences in startup overhead, e.g., due to the need to acquire a GPU, it was measured as the runtime on a trivial instance and subtracted from every runtime. The benchmark score was computed from the points obtained as discussed above by normalizing with the maximum number of obtained points. Consequently, the tool with the most points was assigned a score of 100%. The total competition score was simply the sum of the per benchmark scores, corresponding to equal weighting.

**Results** In 2021, 12 teams participated in the competition.  $\alpha$ - $\beta$ -CROWN won first place, followed by VeriNet in second, and ERAN/OVAL in third, depending on the overhead measurement and voting scheme used to determine result-correctness. Except for VeriNet, they all used the GPU instance.

### 4.3 VNN-COMP 2022

In the most recent iteration of VNN-COMP<sup>4</sup> [36], the evaluation was fully automated, allowing the number of benchmarks to be increased.

<sup>4</sup> <https://sites.google.com/view/vnn2022>.

#### 4.3.1 Benchmarks

In 2022, each participating team could submit or endorse up to two benchmarks, allowing industry practitioners to propose benchmarks without entering a tool. Each benchmark had a total timeout of between three and six hours, with randomization of instances being mandatory this year. Tool tuning was still permitted on a per benchmark level and, in practice, also per network using the network's statistics.

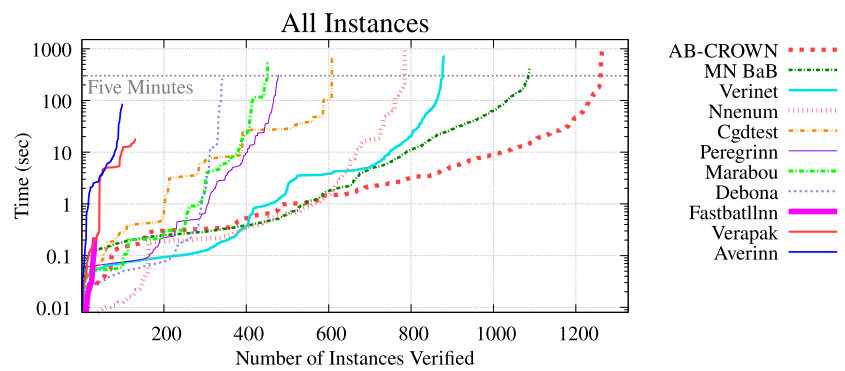
The submitted benchmarks can be grouped into the following categories:

- Fully connected networks with ReLU activations – three benchmarks, based on reinforcement tasks and MNIST.
- Fully connected networks in TLL format [14] – one benchmark.
- Large networks with sparse matrices – one benchmark, based on database indexing and cardinality estimation.
- Convolutional networks – three benchmarks, based on CIFAR10.
- Residual networks – two benchmarks, based on CIFAR10, CIFAR100, and TinyImageNet.
- Complex U-Net networks with average-pooling and softmax – one benchmark based on image segmentation.

#### 4.3.2 Evaluation

Similar to the previous year, teams could choose between a range of AWS instance types (see Table 1) providing a CPU, GPU, or mixed focus. Except for the much weaker t2.large instance, all instances were priced at around three dollars per hour. In contrast to 2021, when organizers had to manually execute installation scripts and debug with the participants, an automated submission and testing pipeline was set up. Teams could submit their benchmarks and tools via a web interface by specifying a git repository, commit hash, and post-installation script (enabling, e.g., the acquisition of licenses). This triggered a new AWS instance to be spawned where all installation scripts were executed. If the installation succeeded, the tool was automatically evaluated on a previously selected set of benchmarks before the instance was terminated again. To enable debugging by the participants, all outputs were logged and made accessible live via the submission website, allowing them to monitor the progress.

**Fig. 1** Cactus plot for all tools in the VNN-COMP 2022 across all benchmarks



**Table 3** Points per instance in 2022

Ground truth	Returned result		
	SAT	UNSAT	Other
SAT	+10	-100	0
UNSAT	-100	+10	0

This automation allowed each team to perform as many tests as necessary without the need to wait for feedback from the organizers. Furthermore, teams could test on the same AWS instances used during final evaluation without having to pay for their usage, with the costs kindly covered by the SRI Lab of ETH Zurich.

**Scoring** Unlike during the VNN-COMP 2021, SAT instances were not divided into simple and complex for scoring purposes, leading to 10 points being awarded for all correct results (see Table 3). Further, instead of relying on a voting scheme to determine the ground truth in the presence of dissent among tools, the burden of proof was placed on the tool reporting SAT, requiring them to provide a concrete counter-example. If no valid counter-example was provided, the corresponding tool was judged to be incorrect and awarded the 100 point penalty.

**Results** Out of the eleven participating teams,  $\alpha$ - $\beta$ -CROWN placed first, MN-BaB second, and VeriNet third. For a comparison of all participating tools across all benchmarks, see Fig. 1.

## 5 Comparison across the years

In Table 4, we list all tools participating in any iteration of the VNN-COMP and refer the interested reader to the corresponding VNN-COMP report for a short description of the tools. In Table 5, we compare the scope of the competition across the last three years. As can be seen, the number, variety, complexity, and scale of benchmarks increased with

every iteration. Starting with 5 benchmarks covering simple fully connected (FC) and convolutional (Conv) networks in 2020, the 2022 competition saw 12 benchmarks including a range of complex residual and U-Net architectures with up to 140 million parameters. Further, we believe that the increasing number of registered tools clearly shows that the interest in both the field in general and the competition in particular is growing year by year. However, the large and increasing discrepancy between registered and submitted tools might indicate that many teams feel like they are not able to invest the significant effort required to support not only the standardized network and specification formats, but also the wide variety of different benchmarks. As tools are ranked by their total score, with each benchmark providing a score of up to 100%, the final ranking is biased towards tools that support all benchmarks. While we believe that this is a valuable incentive for tool developers to develop methods that can be easily applied to new problems, it might be daunting for new teams to implement all necessary features, deterring them from participating at all.

**Successful trends** While all teams started out using only CPUs in 2020, only one of the top four teams relied solely on CPUs in 2021, and all top three teams chose GPU instances in 2022. This transition enabled both the more efficient evaluation of simple bound propagation methods such as DeepPoly [50], CROWN [67], and IBP [19], and approximate solutions of the linear programming (LP) problems arising during verification [15, 60, 63]. Similarly, the top two teams in 2021 and all top three teams in 2022 relied on a branch-and-bound (BaB) based approach, recursively breaking down the verification problem into easier subproblems until it becomes solvable, thus effectively enabling the use of GPUs to solve tighter mixed integer linear programming (MILP) encodings of the verification problem [11, 15, 60, 66]. Both top two teams in the most recent iteration combined this approach with additional multi-neuron [15] and solver-generated cutting plane constraints [66], first introduced by the 3rd place ERAN in 2021 [38]. We thus conclude that successful tools leverage hardware accelerators such as GPUs to efficiently handle tight (MI)LP encodings of the verification problem.



**Table 4** Participating tools

Tool	Organization	Participation, place			References
		2020 [2]	2021 [5]	2022 [36]	
$\alpha$ - $\beta$ -CROWN	Carnegie Mellon, Northeastern, Columbia, UCLA	✗	✓(1/12)	✓(1/11)	[62, 63, 68]
AveriNN	Kansas State University	✗	✗	✓(11/11)	N/A
CGDTest	University of Waterloo	✗	✗	✓(5/11)	N/A
Debona	RWTH Aachen University	✗	✓(6/12)	✓(8/11)	[9]
DNNF	University of Virginia	✗	✓(12/12)	✗	[47]
ERAN	ETH Zurich, UIUC	✓	✓(3/12)	✗	[38, 44, 48–51]
FastBatLLNN	University of California	✗	✗	✓(9/11)	N/A
Marabou	Hebrew University of Jerusalem, Stanford University, Amazon Web Services, NRI Secure	✗	✓(5/12)	✓(7/11)	[26]
MIPVerify	Massachusetts Institute of Technology	✓	✗	✗	[54]
MN-BAB	ETH Zurich	✗	✗	✓(2/11)	[15]
nenum	Stony Brook University	✓	✓(8/12)	✓(4/11)	[4]
NNV	Vanderbilt University	✓	✓(9/12)	✗	[55–58, 61]
NV.jl	Carnegie Mellon, Northeastern	✗	✓(10/12)	✗	[29, 30]
Oval	University of Oxford	✓	✓(3/12)	✗	[10–12, 33, 40–42]
PeregrinNN	University of California	✓	✗	✓(6/11)	[27]
RPM	Stanford	✗	✓(11/12)	✗	[59]
Venus	Imperial College London	✓	✓(7/12)	✗	[8, 28]
VeraPak	Utah State University	✗	✗	✓(10/11)	N/A
VeriNet	Imperial College London	✓	✓(2/12)	✓(3/11)	[21, 22]

**Table 5** Comparison across years

	2020	2021	2022
Tools registered	N/A	15	18
Tools submitted	8	13	11
Benchmarks submitted	5	8 (+1 unscored)	12 (+1 unscored)
Max. network depth	8	18	27
Max. network parameters	855,600	42,059,431 (sparse)	138,356,520
Activation functions	ReLU, tanh, sigmoid	ReLU, sigmoid, MaxPool, AveragePool	ReLU, sigmoid, MaxPool
Layer types	Fully Connected, Conv	Fully Connected, Conv, Residual	Fully Connected, Conv, Residual, BatchNorm
Applications	Image Recognition, Control	Image Recognition, Control, Database Indexing	Image Recognition, Control, Database Indexing, Cardinality Estimation
Mean #benchmarks/tool	3.0 (min 2, max 5)	5.5 (min 1, max 9)	7.3 (min 1, max 13)

## 6 Outlook

Below we discuss considerations that could enable future iterations of the VNN-COMP to serve its goals and the community, discussed in Sect. 3, even better.

### 6.1 Tracking year-on-year progress

While we believe VNN-COMP already provides reasonable mechanisms for comparing the tools submitted in every it-

eration, the changing benchmarks and tools make it hard to track the year-on-year progress of the field as a whole. Because some tools are heavily optimized for the specific benchmarks of that year's competition, simply evaluating them on the benchmarks of previous (or future) years (even if they support them) does not yield a meaningful progress metric. While one benchmark from the inaugural competition was included as an unscored extra benchmark in the two following iterations (*cifar2020*), only few unsolved instances remain, making it a very insensitive measure for

further improvements. While including all benchmarks from previous years in the (scored) benchmark selection would place an undue burden on participants, choosing one particularly challenging, representative, and interesting benchmark every year to be included as a (scored) extra benchmark in future iterations might be a good compromise. Additionally, a more restrictive stance on tool tuning could enable a much more representative evaluation of new tools on old benchmarks.

## 6.2 Tool tuning

Many of the most successful tools do not employ a single verification strategy, but a whole portfolio of different modes, all coming with different hyperparameters. Depending on their choice, tool performance can vary significantly, making it essential for practitioners to get their choice right when applying these tools to new problems. However, this can be highly challenging given the large number of parameters and their complex interactions, especially without in-depth knowledge of the tool.

For VNN-COMP, tuning tools was allowed explicitly on a per-benchmark basis and implicitly on a per-network basis, enabling teams to showcase the maximum performance of their tools. However, for future iterations, it might be interesting to restrict tuning for some or all benchmarks to encourage authors to develop autotuning strategies, making the adaption of their tools to new problems much easier. This could, for example, be implemented by not only generating random specifications, but also random networks.

## 6.3 Batch processing

Every VNN-COMP benchmark consists of a set of instances that, while typically related, are evaluated in isolation, with the tool being terminated in between. Unfortunately, this means that any startup overhead such as acquiring a GPU or preprocessing the considered network is incurred for every instance. This is in contrast to most practical settings where a large number of input-output specifications are considered for the same network. This discrepancy is accounted for by measuring and subtracting this overhead from each individual runtime. However, not only is this overhead measurement process flawed and introduces noise, but it can also dominate the evaluation time for easy instances.

In future iterations, tools could be provided with a whole batch of properties at once to more closely relate to their typical application. Further, currently, timeouts are defined per instance, making a strategy of always attempting verification until timeout is optimal. However, in a practical setting, recognizing instances where verification is likely to fail and stopping early can significantly increase a method's throughput and thus utility. Switching to per benchmark timeouts for

the VNN-COMP would incentivize the development of effective heuristics towards this goal. Furthermore, tools could benefit from proof-sharing approaches [16], where verified sub-problems from one instance are reused for following instances.

## 6.4 Continuous competition

In addition to a yearly VNN-COMP, tool submissions for the most recent benchmark set could be accepted on a rolling basis, made possible by the automated submission and evaluation process introduced this year. This would transform the competition from a yearly snapshot of the current research to a centralized repository of the state-of-the-art, updating as teams submit new methods that they publish. However, if not implemented with great care, this would enable tools to be tuned on the evaluation instances before submission, leading to a skewed comparison. Further, the question of funding the required cloud compute remains open.

## 6.5 Soundness evaluation

An inherent requirement for neural network verifiers is that they be sound, i.e., that they never claim a safety property holds, when in fact it does not. However, assessing soundness is difficult as the ground truth for VNN-COMP problem instances is generally only known if it was shown to be SAT with a valid counter-example. This is particularly problematic when no instances in a benchmark are SAT and thus returning UNSAT for every instance immediately can not be demonstrated to be unsound. Requiring a certain portion of instances of every benchmark to be SAT (in expectation), could alleviate this issue. An interesting alternative avenue to tackle this challenge is proof generation [39]. An extra category could be introduced where tools are additionally required or incentivized to provide a verifiable proof if they claim a property is UNSAT.

While big soundness bugs are rare, few or none of the submitted tools are floating point sound, i.e., even tools that would be sound using exact arithmetic might become unsound due to imprecisions introduced by floating-point arithmetic. This is particularly pronounced if tools choose to use single precision computations for performance reasons. The sensitivity of different tools to such issues could be evaluated on a benchmark specifically designed to uncover floating point soundness issues.

## 6.6 Other competition modes

A dedicated falsifier category could be added to encourage teams to develop and submit stronger attacks, going beyond the standard adversarial attacks. Further, a meta-solver category could be added to investigate whether approaches that



heuristically pick from a range of methods, successful in other domains [64], can significantly outperform individual tools. However, it would need to be ensured that these tools provide sufficient value over individual submissions, which already combine different verification strategies.

## 6.7 Promote common tool development

Parsing large and complex VNN-LIB files or converting ONNX files to other common formats can be time-consuming to implement. While many teams implemented their own tools to this end, available, open-source tools for the parsing of VNN-LIB files [1] and the optimization of ONNX files (DNNV [46]) should be highlighted and their continued development encouraged.

## 6.8 Remaining challenges

We can broadly identify four groups of challenges in neural network verification:

- Verifying relatively small but only weakly regularized networks, which requires an extraordinarily precise analysis, can still be intractable with current methods.
- Scaling precise methods to medium-sized networks (e.g., small ResNets) and datasets (e.g., Cifar100 or TinyImageNet) with a large number of neurons is challenging, as the cost of branch-and-bound based algorithms scales exponentially with the required split depth, making branching decisions both harder and more important.
- Scaling verification to large networks (e.g., VGG-Net 16) and datasets (e.g., ImageNet) in the presence of dense input specifications requires particularly memory-efficient implementations due to a large number of neurons.
- Verification outside of the classification setting is under-explored leading to a lack of established approaches for e.g., image segmentation or object detection.

Orthogonally, the training of certifiably robust networks remains an open problem. Despite significant progress over recent years [6, 19, 35, 37, 43, 45, 65], networks trained specifically to exhibit provable robustness guarantees still suffer from severely degraded standard accuracy. Therefore, most benchmarks considered in the VNN-COMP are based on networks trained without consideration for later certification. More broadly in the community, readers may also be interested in the International Competition on Verifying Continuous and Hybrid Systems (ARCH-COMP)<sup>5</sup> category on Artificial Intelligence and Neural Network Control Systems (AINNCS), which has been held annually since 2019 [31, 32], and considers neural network verification in closed-loop systems.

<sup>5</sup> <https://cps-vo.org/group/ARCH/FriendlyCompetition>.

## 7 Advice for participants

In this section, we provide some guidance for teams that are interested in the VNN-COMP but have not participated yet. Note that these are neither rules nor requirements.

### 7.1 For benchmark authors

The VNN-COMP intends to highlight areas where neural network verification can be successfully applied, and to showcase interesting differences between the participating tools. Thus, ideally, tasks are not so hard that none of the instances can be solved by any participant, but also not so easy that every tool can solve all of them. For benchmarks related to real-world applications, we recommend including a detailed description of the background, to highlight the benchmark's relevance and the characteristics of the verification problem, e.g., sparseness of the input or some network layers. Further questions and requests for modifications should be expected while tool authors work on supporting the proposed benchmark.

### 7.2 For tool authors

We recommend teams reference past benchmarks to test their tool before the new benchmarks are submitted. Given the ever-increasing diversity of submitted benchmarks, it may not be feasible to support all benchmarks from the get-go. If this is the case, we recommend focusing on the fully connected and convolutional ReLU networks, which in the past have covered a wide range of benchmarks, while minimizing implementation effort. Some operations, e.g., max-pooling can also be simplified to multiple ReLU layers using tools such as DNNV [46]. Further, we recommend extensive testing against adversarial attacks to minimize the chance for soundness errors. For tools that are designed for very specific problems, we also want to encourage authors to submit a relevant benchmark highlighting this specialization. Finally, we recommend reading publications associated with the well-performing tools (see Table 4) to gain a better understanding of the techniques used by successful teams.

## 8 Conclusions

In this report, we summarize the main processes and results of the three VNN-COMP held so far from 2020 to 2022. We highlight the growing interest in the field, expressed in an increasing number of registered teams and considered benchmarks, including some submitted by industry. Further, we observe that every year, the size and complexity not only of the considered networks, but also specifications grew, driving and exemplifying progress in the field. Finally, we

highlight the increase in accessibility of verification methods resulting from the standardized input and output formats and the automated installation and evaluation process required for participation in VNN-COMP.

**Acknowledgements** This material is based upon work supported by the Air Force Office of Scientific Research and the Office of Naval Research under award numbers FA9550-19-1-0288, FA9550-21-1-0121, FA9550-22-1-0019, FA9550-22-1-0450, and N00014-22-1-2156, as well the Defense Advanced Research Projects Agency (DARPA) Assured Autonomy program through contract number FA8750-18-C-0089, and the National Science Foundation (NSF) under grants 1911017, 2028001, 2220401, and 2220426. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Air Force, the United States Navy, DARPA, or NSF.

**Funding Note** Open Access funding enabled and organized by Projekt DEAL.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

1. Simple Adversarial Generator. [https://github.com/stanleybak/simple\\_adversarial\\_generator](https://github.com/stanleybak/simple_adversarial_generator). Accessed: 2022-09-13
2. VNN-COMP2020 report. <https://www.overleaf.com/project/5f0c85e8d15dc10001749fa9>. Accessed: 2022-08-28
3. Bai, J., Lu, F., Zhang, K., et al.: Onnx: open neural network exchange (2019). <https://github.com/onnx/onnx>
4. Bak, S.: Execution-guided overapproximation (ego) for improving scalability of neural network verification (2020)
5. Bak, S., Liu, C., Johnson, T.: The second international verification of neural networks competition (VNN-COMP 2021): summary and results (2021). <https://doi.org/10.48550/ARXIV.2109.00498>
6. Balunovic, M., Vechev, M.T.: Adversarial training and provable defenses: bridging the gap. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020 (2020). <https://openreview.net/forum?id=SJxSDxrKDr>
7. Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L.D., Monfort, M., Muller, U., Zhang, J., Zhang, X., Zhao, J., Zieba, K.: End to end learning for self-driving cars (2016). <https://doi.org/10.48550/ARXIV.1604.07316>
8. Botoeva, E., Kouvaros, P., Kronqvist, J., Lomuscio, A., Misener, R.: Efficient verification of neural networks via dependency analysis. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI20). AAAI Press, Menlo Park (2020)
9. Brix, C., Noll, T.: Debona: decoupled boundary network analysis for tighter bounds and faster adversarial robustness proofs. CoRR (2020). [arXiv:2006.09040](https://arxiv.org/abs/2006.09040) [abs]
10. Bunel, R., De Palma, A., Desmaison, A., Dvijotham, K., Kohli, P., Torr, P.H., Kumar, M.P.: Lagrangian decomposition for neural network verification. In: Conference on Uncertainty in Artificial Intelligence (2020)
11. Bunel, R., Lu, J., Turkaslan, I., Kohli, P., Torr, P., Kumar, M.P.: Branch and bound for piecewise linear neural network verification. *J. Mach. Learn. Res.* **21**, 1574–1612 (2020)
12. Bunel, R., Turkaslan, I., Torr, P.H.S., Kohli, P., Mudigonda, P.K.: A unified view of piecewise linear neural network verification. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*, vol. 31, pp. 4795–4804. Curran Associates, Red Hook (2018). <https://proceedings.neurips.cc/paper/2018/hash/be53d253d6bc3258a8160556dda3e9b2-Abstract.html>
13. Ehlers, R.: Formal verification of piece-wise linear feed-forward neural networks. In: International Symposium on Automated Technology for Verification and Analysis, pp. 269–286 (2017). [https://doi.org/10.1007/978-3-319-68167-2\\_19](https://doi.org/10.1007/978-3-319-68167-2_19)
14. Ferlez, J., Shoukry, Y.: ARen: assured ReLU NN architecture for model predictive control of LTI systems. In: Proceedings of the 23rd International Conference on Hybrid Systems: Computation and Control, HSCC '20. ACM, New York (2020). <https://doi.org/10.1145/3365365.3382213>
15. Ferrari, C., Müller, M.N., Jovanovic, N., Vechev, M.T.: Complete verification via multi-neuron relaxation guided branch-and-bound. In: 10th International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25–29, 2022 (2022). [https://openreview.net/forum?id=l\\_amHf1oaK](https://openreview.net/forum?id=l_amHf1oaK)
16. Fischer, M., Sprecher, C., Dimitrov, D.I., Singh, G., Vechev, M.T.: Shared certificates for neural network verification. In: Shoham, S., Vizel, Y. (eds.) *Computer Aided Verification – 34th International Conference, CAV 2022, Proceedings, Part I*, Haifa, Israel, August 7–10, 2022. Lecture Notes in Computer Science, vol. 13371, pp. 127–148. Springer, Berlin (2022). [https://doi.org/10.1007/978-3-031-13185-1\\_7](https://doi.org/10.1007/978-3-031-13185-1_7)
17. Gehr, T., Mirman, M., Drachler-Cohen, D., Tsankov, P., Chaudhuri, S., Vechev, M.T.: AI2: safety and robustness certification of neural networks with abstract interpretation. In: 2018 IEEE Symposium on Security and Privacy, SP 2018, Proceedings, San Francisco, California, USA, 21–23 May 2018, pp. 3–18. IEEE Comput. Soc., Los Alamitos (2018). <https://doi.org/10.1109/SP.2018.00058>
18. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings, San Diego, CA, USA, May 7–9, 2015 (2015). <http://arxiv.org/abs/1412.6572>
19. Goyal, S., Dvijotham, K., Stanforth, R., Bunel, R., Qin, C., Uesato, J., Arandjelovic, R., Mann, T.A., Kohli, P.: On the effectiveness of interval bound propagation for training verifiably robust models. CoRR (2018). [arXiv:1810.12715](https://arxiv.org/abs/1810.12715) [abs]
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016). <https://doi.org/10.1109/CVPR.2016.90>
21. Henriksen, P., Lomuscio, A.: Efficient neural network verification via adaptive refinement and adversarial search. In: Proceedings of the 24th European Conference on Artificial Intelligence (ECAI20) (2020)
22. Henriksen, P., Lomuscio, A.: Deepsplit: an efficient splitting method for neural network verification via indirect effect analysis. In: Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI21) (2021). <https://doi.org/10.24963/ijcai.2021/351>
23. Huang, X., Kwiatkowska, M., Wang, S., Wu, M.: Safety verification of deep neural networks. In: Majumdar, R., Kunčák, V. (eds.) *Computer Aided Verification*, pp. 3–29. Springer, Cham (2017)

24. Julian, K.D., Lopez, J., Brush, J.S., Owen, M.P., Kochenderfer, M.J.: Policy compression for aircraft collision avoidance systems. In: 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC), pp. 1–10 (2016). <https://doi.org/10.1109/DASC.2016.7778091>
25. Katz, G., Barrett, C., Dill, D.L., Julian, K., Kochenderfer, M.J.: ReLuplex: An efficient SMT solver for verifying deep neural networks. In: Majumdar, R., Kunčák, V. (eds.) *Computer Aided Verification*, pp. 97–117. Springer, Cham (2017)
26. Katz, G., Huang, D.A., Ibeling, D., Julian, K., Lazarus, C., Lim, R., Shah, P., Thakoor, S., Wu, H., Zeljić, A., et al.: The Marabou framework for verification and analysis of deep neural networks. In: *International Conference on Computer Aided Verification*, pp. 443–452. Springer, Berlin (2019)
27. Khedr, H., Ferlez, J., Shoukry, Y.: Effective formal verification of neural networks using the geometry of linear regions. *arXiv preprint* (2020). [arXiv:2006.10864](https://arxiv.org/abs/2006.10864)
28. Kouvaros, P., Lomuscio, A.: Towards scalable complete verification of ReLU neural networks via dependency-based branching. In: *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI21)* (2021). <https://doi.org/10.24963/ijcai.2021/364>
29. Liu, C., Arnon, T., Lazarus, C., Kochenderfer, M.J.: Neuralverification.jl: algorithms for verifying deep neural networks. In: *ICLR 2019 Debugging Machine Learning Models Workshop* (2019). [https://debug-ml-iclr2019.github.io/cameraready/DebugML-19\\_paper\\_22.pdf](https://debug-ml-iclr2019.github.io/cameraready/DebugML-19_paper_22.pdf)
30. Liu, C., Arnon, T., Lazarus, C., Strong, C., Barrett, C., Kochenderfer, M.J.: Algorithms for verifying deep neural networks. *Found. Trends Optim.* **4**(3–4), 244–404 (2021). <https://doi.org/10.1561/24000000035>
31. Lopez, D.M., Althoff, M., Benet, L., Chen, X., Fan, J., Forests, M., Huang, C., Johnson, T.T., Ladner, T., Li, W., Schilling, C., Zhu, Q.: Arch-comp22 category report: artificial intelligence and neural network control systems (AINNCS) for continuous and hybrid systems plants. In: *Proceedings of 9th International Workshop on Applied Verification of Continuous and Hybrid Systems (ARCH22)*. EPiC Series in Computing, vol. 90, pp. 142–184 (2022). <https://doi.org/10.29007/wfgr>
32. Lopez, D.M., Musau, P., Tran, H.D., Dutta, S., Carpenter, T.J., Ivanov, R., Johnson, T.T.: Arch-comp19 category report: artificial intelligence and neural network control systems (ainnecs) for continuous and hybrid systems plants. In: *Proceedings of 6th International Workshop on Applied Verification of Continuous and Hybrid Systems*. EPiC Series in Computing, vol. 61, pp. 103–119 (2019). <https://doi.org/10.29007/rgv8>
33. Lu, J., Kumar, M.P.: Neural network branching for neural network verification. In: *International Conference on Learning Representations* (2020)
34. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: *6th International Conference on Learning Representations, ICLR 2018, Conference Track Proceedings*, Vancouver, BC, Canada, April 30–May 3, 2018 (2018). <https://openreview.net/forum?id=rJzIBfZAb>
35. Mirman, M., Gehr, T., Vechev, M.T.: Differentiable abstract interpretation for provably robust neural networks. In: *Dy, J.G., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10–15, 2018. Proceedings of Machine Learning Research*, vol. 80, pp. 3575–3583 (2018). <http://proceedings.mlr.press/v80/mirman18b.html>
36. Müller, M.N., Brix, C., Bak, S., Liu, C., Johnson, T.T.: The third international verification of neural networks competition (VNN-COMP 2022): summary and results (2022). <https://doi.org/10.48550/arXiv.2212.10376>
37. Müller, M.N., Eckert, F., Fischer, M., Vechev, M.T.: Certified training: small boxes are all you need. *CoRR* (2022). <https://doi.org/10.48550/arXiv.2210.04871>
38. Müller, M.N., Makarchuk, G., Singh, G., Püschel, M., Vechev, M.: Prima: precise and general neural network certification via multi-neuron convex relaxations. *arXiv preprint* (2021). [arXiv:2103.03638](https://arxiv.org/abs/2103.03638)
39. Isac, O., Barrett, C., Zhang, M., Katz, G.: Neural network verification with proof production. In: *22nd International Conference on Formal Methods in Computer-Aided Design (FMCAD)* (2022)
40. De Palma, A., Behl, H.S., Bunel, R., Torr, P.H.S., Kumar, M.P.: Scaling the convex barrier with active sets. In: *9th International Conference on Learning Representations, ICLR 2021, Conference Track Proceedings*, May 3–7, 2021 (2021). <https://openreview.net/forum?id=uQfOy7LrITR>
41. De Palma, A., Behl, H.S., Bunel, R., Torr, P.H.S., Kumar, M.P.: Scaling the convex barrier with sparse dual algorithms. *CoRR* (2021). <https://doi.org/10.48550/arXiv.2101.05844>
42. De Palma, A., Bunel, R., Desmaison, Alban., Dvijotham, K., Kohli, P., Torr, P.H.S., Kumar, M.P.: Improved branch and bound for neural network verification via lagrangian decomposition. *CoRR* (2021). <https://doi.org/10.48550/arXiv.2104.06718>
43. De Palma, A., Bunel, R., Dvijotham, K., Kumar, M.P., Stanforth, R.: IBP regularization for verified adversarial robustness via branch-and-bound. (2022). <https://doi.org/10.48550/arXiv.2206.14772>
44. Serre, F., Müller, C., Singh, G., Püschel, M., Vechev, M.: Scaling polyhedral neural network verification on GPUs. In: *Proc. Machine Learning and Systems (MLSys)* (2021)
45. Shi, Z., Wang, Y., Zhang, H., Yi, J., Hsieh, C.: Fast certified robust training with short warmup. In: *Ranzato, M., Beygelzimer, A., Dauphin, Y.N., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, Virtual, December 6–14, 2021*, pp. 18335–18349 (2021). <https://proceedings.neurips.cc/paper/2021/hash/988f9153ac4fd966ea302dd9ab9bae15-Abstract.html>
46. Shriver, D., Elbaum, S., Dwyer, M.B.: DNNV: a framework for deep neural network verification. In: *Silva, A., Leino, K.R.M. (eds.) Computer Aided Verification*, pp. 137–150. Springer, Cham (2021)
47. Shriver, D., Elbaum, S.G., Dwyer, M.B.: Reducing DNN properties to enable falsification with adversarial attacks. In: *43rd IEEE/ACM International Conference on Software Engineering, ICSE 2021, Madrid, Spain, 22–30 May 2021*, pp. 275–287. IEEE (2021). <https://doi.org/10.1109/ICSE43902.2021.00036>
48. Singh, G., Ganvir, R., Püschel, M., Vechev, M.: Beyond the single neuron convex barrier for neural network certification. In: *Advances in Neural Information Processing Systems*, vol. 32, pp. 15098–15109. Curran Associates, Red Hook (2019)
49. Singh, G., Gehr, T., Mirman, M., Püschel, M., Vechev, M.: Fast and effective robustness certification. In: *Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems*, vol. 31, pp. 10802–10813. Curran Associates, Red Hook (2018). <http://papers.nips.cc/paper/8278-fast-and-effective-robustness-certification.pdf>
50. Singh, G., Gehr, T., Püschel, M., Vechev, M.: An abstract domain for certifying neural networks. *Proc. ACM Program. Lang.* **3**(POPL), 41:1–41:30 (2019)
51. Singh, G., Gehr, T., Püschel, M., Vechev, M.: Boosting robustness certification of neural networks. In: *Proc. International Conference on Learning Representations (ICLR)* (2019)
52. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I.J., Fergus, R.: Intriguing properties of neural networks. In: *Bengio, Y., LeCun, Y. (eds.) 2nd International Conference on Learning Representations, ICLR 2014, Conference Track*

- Proceedings, Banff, AB, Canada, April 14–16, 2014 (2014). <http://arxiv.org/abs/1312.6199>
53. Tacchella, A., Pulina, L., Guidotti, D., Demarchi, S.: The verification of neural networks library (VNN-LIB) (2019). <https://www.vnnlib.org>
  54. Tjeng, V., Xiao, K.Y., Tedrake, R.: Evaluating robustness of neural networks with mixed integer programming. In: ICLR (2019)
  55. Tran, H.D., Bak, S., Xiang, W., Johnson, T.T.: Verification of deep convolutional neural networks using imagestars. In: 32nd International Conference on Computer-Aided Verification (CAV). Springer, Berlin (2020)
  56. Tran, H.D., Musau, P., Lopez, D.M., Yang, X., Nguyen, L.V., Xiang, W., Johnson, T.T.: Parallelizable reachability analysis algorithms for feed-forward neural networks. In: Proceedings of the 7th International Workshop on Formal Methods in Software Engineering (FormalSE'19), FormalSE '19, pp. 31–40. IEEE Press, Piscataway (2019). <https://doi.org/10.1109/FormalSE.2019.00012>
  57. Tran, H.D., Musau, P., Lopez, D.M., Yang, X., Nguyen, L.V., Xiang, W., Johnson, T.T.: Star-based reachability analysis for deep neural networks. In: 23rd International Symposium on Formal Methods (FM'19). Springer, Berlin (2019)
  58. Tran, H.D., Yang, X., Lopez, D.M., Musau, P., Nguyen, L.V., Xiang, W., Bak, S., Johnson, T.T.: NNV: the neural network verification tool for deep neural networks and learning-enabled cyber-physical systems. In: 32nd International Conference on Computer-Aided Verification (CAV) (2020)
  59. Vincent, J.A., Schwager, M.: Reachable polyhedral marching (RPM): a safety verification algorithm for robotic systems with deep neural network components (2021)
  60. Wang, S., Zhang, H., Xu, K., Lin, X., Jana, S., Hsieh, C.J., Kolter, Z.: Beta-CROWN: efficient bound propagation with per-neuron split constraints for complete and incomplete neural network verification. arXiv preprint (2021). [arXiv:2103.06624](https://arxiv.org/abs/2103.06624)
  61. Xiang, W., Tran, H., Johnson, T.T.: Output reachable set estimation and verification for multilayer neural networks. IEEE Trans. Neural Netw. Learn. Syst. **29**(11), 5777–5783 (2018)
  62. Xu, K., Shi, Z., Zhang, H., Wang, Y., Chang, K.W., Huang, M., Kailkhura, B., Lin, X., Hsieh, C.J.: Automatic perturbation analysis for scalable certified robustness and beyond. In: Advances in Neural Information Processing Systems, vol. 33 (2020)
  63. Xu, K., Zhang, H., Wang, S., Wang, Y., Jana, S., Lin, X., Hsieh, C.J.: Fast and complete: enabling complete neural network verification with rapid and massively parallel incomplete verifiers. In: International Conference on Learning Representations (2021). <https://openreview.net/forum?id=nVZtXBI6LNn>
  64. Xu, L., Hutter, F., Hoos, H.H., Leyton-Brown, K.: SATzilla: portfolio-based algorithm selection for SAT. J. Artif. Intell. Res. **32**(1), 565–606 (2008)
  65. Zhang, H., Chen, H., Xiao, C., Goyal, S., Stanforth, R., Li, B., Boning, D.S., Hsieh, C.: Towards stable and efficient training of verifiably robust neural networks. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020 (2020). <https://openreview.net/forum?id=Skxuk1rFwB>
  66. Zhang, H., Wang, S., Xu, K., Li, L., Li, B., Jana, S., Hsieh, C., Kolter, J.Z.: General cutting planes for bound-propagation-based neural network verification. CoRR (2022). <https://doi.org/10.48550/arXiv.2208.05740>
  67. Zhang, H., Weng, T., Chen, P., Hsieh, C., Daniel, L.: Efficient neural network robustness certification with general activation functions. In: Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, Montréal, Canada, December 3–8, 2018, pp. 4944–4953 (2018). <https://proceedings.neurips.cc/paper/2018/hash/d04863f100d59b3eb688a11f95b0ae60-Abstract.html>
  68. Zhang, H., Weng, T.W., Chen, P.Y., Hsieh, C.J., Daniel, L.: Efficient neural network robustness certification with general activation functions. Adv. Neural Inf. Process. Syst. **31**, 4939–4948 (2018). <https://arxiv.org/pdf/1811.00866.pdf>
  69. Zhou, C.: Computation of optical flow using a neural network. In: IEEE 1988 International Conference on Neural Networks, vol. 2, pp. 71–78 (1988). <https://doi.org/10.1109/ICNN.1988.23914>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.