
Optimized Symbolic Interval Propagation for Neural Network Verification

Philipp Kern¹ Marko Kleine Büning¹ Carsten Sinz¹

Abstract

Neural networks are increasingly applied in safety critical domains, their verification thus is gaining importance. A large class of recent algorithms for proving input-output relations of feed-forward neural networks are based on linear relaxations and symbolic interval propagation. However, due to variable dependencies, the approximations deteriorate with increasing depth of the network. In this paper we present DPNEURIFYFV, a novel branch-and-bound solver for ReLU networks with low dimensional input-space that is based on symbolic interval propagation with fresh variables and input-splitting. A new heuristic for choosing the fresh variables allows to ameliorate the dependency problem, while our novel splitting heuristic, in combination with several other improvements, speeds up the branch-and-bound procedure. We evaluate our approach on the airborne collision avoidance networks ACAS Xu and demonstrate runtime improvements compared to state-of-the-art tools.

1. Introduction

Neural networks (NNs) are considered state-of-the-art solutions for many machine learning tasks and are increasingly applied in a variety of different areas. However, the complex structure of neurons, layers, and weights often renders their behavior incomprehensible to humans. This can have a severe impact when NNs are applied to safety-critical systems such as airborne collision avoidance (Julian et al., 2016) or self-driving cars (Bojarski et al., 2016). The increased application of NNs in security-critical areas necessitates formal guarantees about their behavior and functionality.

¹Department of Computer Science, Karlsruhe Institute of Technology, Karlsruhe, Germany. Correspondence to: Philipp Kern <philipp.kern@kit.edu>, Marko Kleine Büning <marko.kleinebuening@kit.edu>, Carsten Sinz <carsten.sinz@kit.edu>.

1st Workshop on Formal Verification of Machine Learning, Baltimore, Maryland, USA. Colocated with ICML 2022. Copyright 2022 by the author(s).

In general, the goal is to verify input-output relations of NNs. For this purpose, there are a number of approaches based on SMT (Ehlers, 2017; Huang et al., 2017) or MILP (Katz et al., 2017; 2019) solving as well as on abstract interpretation theory (Müller et al., 2021a;b; Singh et al., 2019a; Tran et al., 2020; Xiang et al., 2018) including symbolic interval propagation (SIP) (Fischer et al., 2019; Gehr et al., 2018; Henriksen & Lomuscio, 2020; Wang et al., 2018a).

Naïve interval propagation is a fast analysis method to propagate lower and upper bounds through a network. However, it suffers from the *dependency problem* and thus can lead to very coarse overapproximations. Although this problem can be weakened by the application of relational domains like zonotopes, error-based symbolic interval propagation or backward substitution of symbolic intervals (Wang et al., 2018b; Singh et al., 2019b), these solutions are either restricted to parallel relaxations of the activation functions or require a quadratic number of backsubstitution passes. While not as effective in reducing the dependency problem, the introduction of fresh variables in SIP (Paulsen et al., 2020) requires only a single forward pass and allows for the use of non-parallel relaxations.

In this paper, we present an approach that verifies input-output relations of feed-forward ReLU-networks with low-dimensional input space via input splitting and optimized symbolic interval propagation with fresh variables.

The main contributions of this work are: (i) We re-examine and optimize the introduction of fresh variables in symbolic interval propagation by presenting heuristics for the amount of variables to introduce per layer and which neurons to prioritize. (ii) We present a novel input splitting heuristic taking into account internal overapproximation (similar to (Henriksen & Lomuscio, 2021) for error-based symbolic interval propagation and neuron splitting) (iii) We improve counterexample generation by utilizing information of the symbolic bounds instead of just sampling the center of the input hyperrectangle.

2. Foundations and Related Work

2.1. Neural Networks.

A feed-forward neural network \mathcal{N} represents a function $f_{\mathcal{N}} : \mathbb{R}^{d_0} \rightarrow \mathbb{R}^{d_L}$ and consists of an input layer, multiple

hidden layers and an output layer (Goodfellow et al., 2016). The l -th layer ($l \geq 1$) comprises d_l neurons \mathbf{n}_l , whose activation values are calculated as an affine function of the values of the previous layer’s neurons (or the inputs, for $l = 0$, identifying \mathbf{n}_0 with the input vector \mathbf{x}), and then fed through a non-linear activation function σ according to $\mathbf{n}_l = \sigma(W_l \mathbf{n}_{l-1} + \mathbf{b}_l)$, where W_l is the weight matrix and \mathbf{b}_l is the bias vector of layer l . Vector $\mathbf{y} = \mathbf{n}_L$ is the output of the NN. In this work, we only consider cases, where σ is the ReLU activation function defined as $\text{ReLU}(x) = \max(0, x)$ and the weights of the NN are fixed.

2.2. Neural Network Verification.

The field of NN verification is concerned with verifying that a NN computing a function $\mathbf{y} = f_{\mathcal{N}}(\mathbf{x})$ always satisfies a property $P(\mathbf{y})$ for all inputs in a set $\mathcal{X} \subseteq \mathbb{R}^{d_0}$. In most of the current literature, P is represented as a set of linear constraints on the output of the NN, while the input is constrained to values within a hyperrectangle $\mathcal{H} = \{\mathbf{x} \in \mathbb{R}^{d_0} \mid \underline{x}_i \leq x_i \leq \bar{x}_i \text{ for all } 0 \leq i < d_0\}$ given by lower and upper bounds, \underline{x}_i and \bar{x}_i , for each dimension.

If P consists of only a single inequality $\mathbf{c}^T \mathbf{y} + b \leq 0$, the problem of NN verification can be rephrased as an optimization problem (Bunel et al., 2020):

$$\max_{\mathbf{x}} \mathbf{c}^T \mathbf{y} + b \quad (1)$$

$$s.t. \mathbf{y} = f_{\mathcal{N}}(\mathbf{x}) \quad (2)$$

$$\underline{x}_i \leq x_i \leq \bar{x}_i, \forall i. \quad (3)$$

If the maximal value v^* of the optimization problem is larger than 0, the property is violated and a concrete input vector \mathbf{x}^* with $\mathbf{c}^T f_{\mathcal{N}}(\mathbf{x}^*) + b > 0$ can be computed as a counterexample to the validity of property P . If we can prove $v^* \leq 0$ or we can find some upper bound v_{ub}^* with $v^* \leq v_{\text{ub}}^* \leq 0$, then the property is guaranteed to hold for network \mathcal{N} and all $x \in \mathcal{H}$.

This formulation can be generalized to properties P consisting of multiple inequalities. We always assume that these are given in the form $\mathbf{c}^T \mathbf{y} + b \leq 0$. Thus, by maximizing the maximum violation for all inequality constraints, we can check whether P holds or not. Due to the piecewise-linear ReLU activations, NN verification may require case splitting, and is in general NP-complete (Katz et al., 2017). Still many approaches have been proposed, for example based on mixed integer linear programming (Cheng et al., 2017; Tjeng et al., 2017), SMT solvers (Ehlers, 2017; Katz et al., 2017), or abstract interpretation based on different set representations (Bak et al., 2020; Henriksen & Lomuscio, 2020; Singh et al., 2018; 2019b; Wang et al., 2018b;a; Zhang et al., 2018).

2.3. Symbolic Interval Propagation

In abstract interpretation, a suitable set representation S_0 encompassing the admissible values for the input set \mathcal{X} is propagated through the layers of the NN until an output set S_L is obtained, which overapproximates the true output set $\{\mathbf{y} = f_{\mathcal{N}}(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}\}$ of the NN. Despite the NP-completeness of NN verification, abstract interpretation approaches allow for efficient verification, if one can easily check whether $P(\mathbf{y})$ is satisfied for all $\mathbf{y} \in S_L$.

In the following paragraphs we want to give an overview over symbolic intervals (Wang et al., 2018b;a), which are one of the possible set representations in abstract interpretation for NNs. In particular, we want to highlight two main factors that affect the quality that can be obtained from symbolic interval propagation: The dependency problem and the overapproximation introduced, when symbolic intervals are propagated through the non-linear activation functions.

In traditional interval arithmetic (Moore, 1966), intervals $[\underline{x}_i, \bar{x}_i]$ with concrete lower and upper bounds $\underline{x}_i \leq x_i \leq \bar{x}_i$ are chosen as enclosing set representations for the true value of x_i . To obtain a valid enclosing interval for operations like addition, say $x + y$, where $x \in [\underline{x}, \bar{x}]$, $y \in [\underline{y}, \bar{y}]$, the interval extension $\hat{+}$ of the addition operation is defined as $[\underline{x}, \bar{x}] \hat{+} [\underline{y}, \bar{y}] = [\underline{x} + \underline{y}, \bar{x} + \bar{y}]$. A similar extension can be defined for subtraction as $[\underline{x}, \bar{x}] \hat{-} [\underline{y}, \bar{y}] = [\underline{x} - \bar{y}, \bar{x} - \underline{y}]$. Although interval arithmetic is cheap, most of the time it computes very conservative overapproximations. If we subtract the interval $[\underline{x}, \bar{x}]$ representing $x \in \mathbb{R}$ from itself, we obtain $[\underline{x}, \bar{x}] - [\underline{x}, \bar{x}] = [\underline{x} - \bar{x}, \bar{x} - \underline{x}]$ instead of the expected interval $[0, 0]$. The reason for this is the *dependency problem*. Information about dependencies between intervals is lost during computation, and different operands representing the same variable can vary independently within their lower and upper bounds. This problem is amplified by chained computations, where the outputs of one stage serve as inputs for the next stage (Stolfi & Figueiredo, 1997). Therefore abstract interpretation for (deep) NNs is particularly affected by the dependency problem.

As a countermeasure, Wang et al. (Wang et al., 2018b;a) introduce symbolic intervals $[lb(\mathbf{x}), ub(\mathbf{x})]$, where possible values of quantities $z \in \mathbb{R}$ are enclosed by pairs of linear functions $lb(\mathbf{x}) = \mathbf{l}^T \mathbf{x} + l_0$ and $ub(\mathbf{x}) = \mathbf{u}^T \mathbf{x} + u_0$ in the input variables \mathbf{x} that z depends upon, instead of concrete bounds. This allows to capture more dependency information. Note that each input variable x_i only occurs once in each bound, which precludes it from attaining its concrete lower and upper bounds simultaneously. For vector-valued quantities $\mathbf{z} \in \mathbb{R}^n$, $\mathbf{l}b(\mathbf{x})$ and $\mathbf{u}b(\mathbf{x})$ are affine mappings, where the \mathbf{l}^T and \mathbf{u}^T form the rows of a coefficient matrix and l_0 and u_0 are vectors.

Propagation of a symbolic interval vector $[\mathbf{l}b(\mathbf{x}), \mathbf{u}b(\mathbf{x})]$

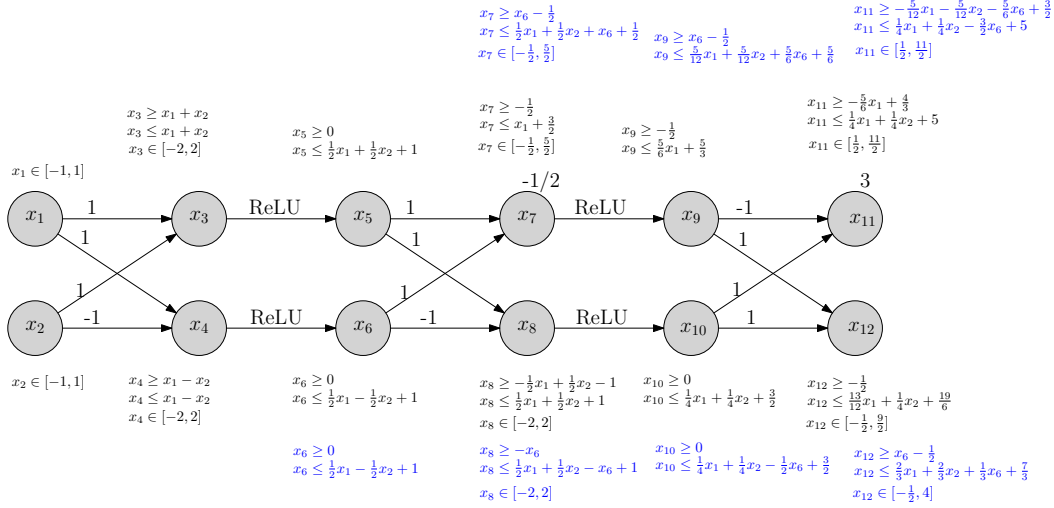


Figure 1: Example for forward propagation of symbolic intervals. Weights are written as edge labels, the bias, if non-zero, is indicated above to the right of the nodes. Symbolic interval propagation is shown in black, while the effects of introducing a fresh variable x_6 for the constraints are shown in blue, resulting in an improved bound for x_{12} .

through affine mappings $\mathbf{y} = W\mathbf{x} + \mathbf{b}$ with constant W and \mathbf{b} results in another symbolic interval defined by $\mathbf{Ib}'(\mathbf{x}) = W^+\mathbf{Ib}(\mathbf{x}) + W^-\mathbf{ub}(\mathbf{x}) + \mathbf{b}$ and $\mathbf{ub}'(\mathbf{x}) = W^+\mathbf{ub}(\mathbf{x}) + W^-\mathbf{Ib}(\mathbf{x}) + \mathbf{b}$, where W^+ and W^- denote the positive and the negative entries of W respectively. As only linear operations are used, no overapproximation is incurred by this propagation (Wang et al., 2018a).

Occurrences of the same input variable with opposite signs can be combined, such that there is only one occurrence of the input variable in each of the output’s symbolic lower and upper bounds, which is set to either its lower or upper bound upon concretization. To illustrate this behaviour, the value of x_7 in Fig. 1 is calculated as $x_7 = x_5 + x_6 - \frac{1}{2}$, where the true values are enclosed by symbolic intervals $x_5 \in [0, \frac{1}{2}x_1 + \frac{1}{2}x_2 + 1]$ and $x_6 \in [0, \frac{1}{2}x_1 - \frac{1}{2}x_2 + 1]$ for $x_1, x_2 \in [-1, 1]$. We obtain the symbolic upper bound of x_7 as

$$\begin{aligned} ub_7(x_1, x_2) &= ub_5(x_1, x_2) + ub_6(x_1, x_2) - \frac{1}{2} \\ &= \left(\frac{1}{2}x_1 + \frac{1}{2}x_2 + 1\right) + \left(\frac{1}{2}x_1 - \frac{1}{2}x_2 + 1\right) - \frac{1}{2} \\ &= x_1 + \frac{3}{2}, \end{aligned}$$

where the opposite signs of x_2 even cancel completely. The concrete upper bound of $\bar{x}_7 = \frac{5}{2}$ can be obtained by setting $x_1 \in [-1, 1]$ to its upper bound since it occurs with positive coefficient in $ub_7(x_1, x_2)$. If we had just used the concrete interval enclosure of $x_5 \in [\underline{x}_5, \bar{x}_5] = [0, 2]$, $x_6 \in [\underline{x}_6, \bar{x}_6] = [0, 2]$, we would have gotten an unnecessarily loose enclosing interval of $x_7 \in [-\frac{1}{2}, \frac{7}{2}]$, as we would have

implicitly used the upper bound \bar{x}_2 in \bar{x}_5 and the lower bound \underline{x}_2 in \bar{x}_6 .

In order to propagate symbolic intervals through the piecewise linear activation function $\text{ReLU}(x)$, for $x \in [lb(\mathbf{x}), ub(\mathbf{x})]$, the symbolic interval is concretized to obtain a concrete interval enclosure $x \in [l, u]$. The symbolic interval enclosing the output of $y = \text{ReLU}(x)$ is then computed as

$$lb'(\mathbf{x}) = \begin{cases} 0 & , u \leq 0 \\ \underline{\text{ReLU}}(lb(\mathbf{x})) & , l < 0 < u \\ lb(\mathbf{x}) & , l \geq 0 \end{cases}$$

and

$$ub'(\mathbf{x}) = \begin{cases} 0 & , u \leq 0 \\ \overline{\text{ReLU}}(ub(\mathbf{x})) & , l < 0 < u \\ ub(\mathbf{x}) & , l \geq 0 \end{cases},$$

where the ReLU function is linearly overapproximated by lower and upper relaxations $\underline{\text{ReLU}}(x)$ and $\overline{\text{ReLU}}(x)$, it is unstable or crossing ($l < 0 < u$), while no overapproximation is necessary if it is fixed-active ($l \geq 0$) or fixed-zero ($u \leq 0$).

While error-based symbolic interval propagation (Henriksen & Lomuscio, 2020) and zonotope propagation (Singh et al., 2018) are restricted to parallel ReLU relaxations, the methods of (Zhang et al., 2018) and (Singh et al., 2019b) use different slopes for the upper relaxation $\overline{\text{ReLU}}(x) =$

$\frac{u}{u-l}x - \frac{ul}{u-l}$ and the lower relaxation

$$\underline{\text{ReLU}}(x) = \begin{cases} 0, & u < |l| \\ x, & u \geq |l| \end{cases}$$

where $\underline{\text{ReLU}}(x)$ is chosen adaptively to reduce the area of the introduced overapproximation.

This ReLU relaxation is also used for analysis of the example NN shown in Fig. 1. Utilizing the fact that $\text{ReLU}(x)$ is monotonic in x , the authors of (Wang et al., 2018b) introduce, what we call separate relaxations. They calculate separate concrete intervals $lb(\mathbf{x}) \in [l_l, l_u]$ and $ub(\mathbf{x}) \in [u_l, u_u]$. The symbolic bounds $lb(\mathbf{x})$ and $ub(\mathbf{x})$ are then separately propagated through the lower parallel relaxation of $\text{ReLU}(x)$ for $x \in [l_l, l_u]$ and the upper parallel relaxation of $\text{ReLU}(x)$ for $x \in [u_l, u_u]$.

Although symbolic interval propagation ensures that *input variables* never attain their lower and upper bound at the same time (assuming they are not equal), when concretizing the symbolic interval, no such guarantees can be given for intermediate neurons. For example, we implicitly use the *upper* relaxation of neuron x_6 in network in Fig. 1 to calculate the symbolic upper bounds $ub_7(\mathbf{x})$ of x_7 , which we use for the calculation of the $ub_9(\mathbf{x})$, which is in turn used for $ub_{12}(\mathbf{x})$. This symbolic upper bound of x_{12} , however is calculated using $ub_{10}(\mathbf{x})$, which relies on $ub_8(\mathbf{x})$, which is calculated using the *lower* relaxation of x_6 .

Zonotope- (Singh et al., 2018) and error-based symbolic interval propagation (Henriksen & Lomuscio, 2020) as well as backsubstitution methods (Zhang et al., 2018; Singh et al., 2019b) can solve this problem. Both methods, however, cannot utilize separate relaxations and while the former methods are restricted to only parallel relaxations, the number of backsubstitution passes required by the latter methods grows quadratically in the number of layers.

In the context of verification of equivalence for NNs, the authors of (Paulsen et al., 2020) introduce symbolic interval propagation with the additional introduction of fresh variables that are used for some of the overapproximated neurons. In order to introduce a fresh variable for a neuron x_{n+1} contained within the symbolic interval $[lb_{n+1}(\mathbf{x}), ub_{n+1}(\mathbf{x})]$ depending on the input $\mathbf{x} \in \mathbb{R}^n$ of the NN, they store its symbolic bounds for later substitution and instead propagate the new symbolic interval $[x_{n+1}, x_{n+1}]$, ensuring that x_{n+1} can never simultaneously operate at its lower and upper symbolic bounds. A symbolic interval $[lb'(\mathbf{x}), ub'(\mathbf{x})]$ only depending on the inputs of the NN, can be recovered from a symbolic interval $[l_0 + \sum_{i=1}^{n+k} l_i x_i, u_0 + \sum_{i=1}^{n+k} u_i x_i]$ with k fresh variables

according to

$$lb'(\mathbf{x}) = l_0 + \sum_{i=1}^n l_i x_i + \sum_{i=n+1}^{n+k} (l_i^+ lb_i(\mathbf{x}) + l_i^- ub_i(\mathbf{x}))$$

$$ub'(\mathbf{x}) = u_0 + \sum_{i=1}^n u_i x_i + \sum_{i=n+1}^{n+k} (u_i^+ ub_i(\mathbf{x}) + u_i^- lb_i(\mathbf{x}))$$

where the $l_i^+ = \max(0, l_i)$, $u_i^+ = \max(0, u_i)$ represent positive and the $l_i^- = \min(0, l_i)$, $u_i^- = \min(0, u_i)$ represent negative coefficients in the lower and upper symbolic bounds respectively. Some dependency information is lost, however, as only symbolic bounds with respect to the input variables are stored in order to avoid expensive chained backsubstitutions. Furthermore, the introduction of fresh variables in later layers impedes the cancellation of earlier variables (which would again increase the dependency problem), because all of their coefficients are set to zero in the new symbolic interval for the new fresh variable. Nevertheless, this approach allows for the use of non-parallel ReLU relaxations and is capable of decreasing the impact of the dependency issue, when fresh variables are introduced for fitting unstable neurons.

For example, introducing a fresh variable for neuron x_6 in Fig. 1, results in an improvement in the concrete upper bound of the output x_{12} . The corresponding symbolic intervals that result from the introduction of x_6 as a fresh variable are shown in blue.

2.4. Branch and Bound

Although symbolic interval propagation provides a relatively cheap way of obtaining bounds on the maximal violation v^* in Equation 3, these bounds are not necessarily sufficient to prove the corresponding property. A branch and bound algorithm (such as the one presented by Bunel et al. (Bunel et al., 2020)) uses a `split` function to repeatedly divide the current problem into smaller sub-problems on which more accurate bounds on the maximal violation can be computed. A split either divides the input domain into subdomains or an unfixed ReLU activation unit into different phases (Bunel et al., 2020).

Different choices for splitting as well as computation of the lower and upper bounds affect the performance of the branch and bound algorithm. For example ZOPE (Strong et al., 2021) uses zonotope propagation to obtain upper bounds on maximization problems over the outputs of NNs and samples the center \mathbf{x}_{mid} of the input hyperrectangle \mathcal{H} to obtain a lower bound $f_{\mathcal{N}}(\mathbf{x}_{mid})$ on the maximal achievable value. For the split operation, Strong et al. (Strong et al., 2021) chose to bisect the input x_i with the largest range $[x_i, \bar{x}_i]$ into two halves.

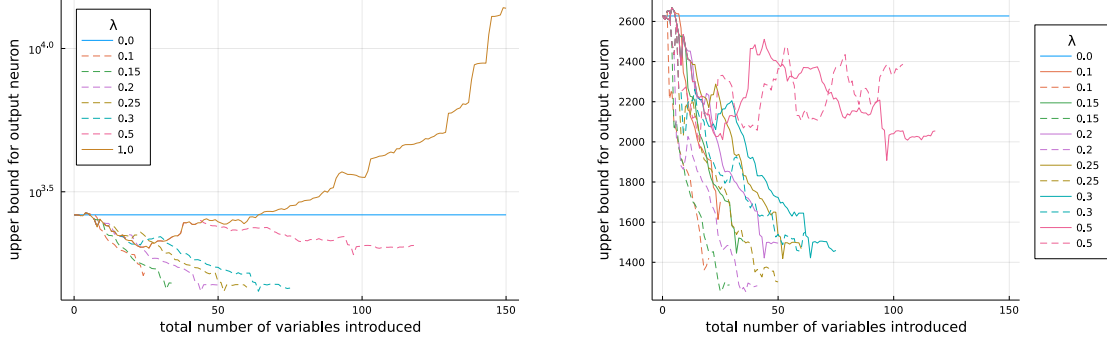


Figure 2: Upper bounds on the first output of ACAS-Xu.1.1 for the input space of property ϕ_1 obtained from a single symbolic forward pass with varying total number of introduced variables. Left: Introducing variables as early as possible for different parameters λ . No variables ($\lambda = 0$) and the approach of (Paulsen et al., 2020) ($\lambda = 1$) are shown as solid lines. Right: For different parameters λ the effect of introducing a number of variables as early as possible is shown as solid lines, the effect of prioritizing neurons with large concrete ranges and ignoring fixed-zero neurons is shown as dashed lines.

3. Our Approach

Our approach is based on a combination and thorough refinement of previously presented methods. We give a brief overview, before presenting the details. In order to reduce the effect of the dependency issue, we utilize symbolic interval propagation with fresh variables (Paulsen et al., 2020) and improve its efficacy by finding heuristics for the amount of variables to introduce per layer and which neurons to prioritize. These optimizations are described in Sec. 3.1.

To keep the overapproximations at unstable neurons as small as possible, we utilize the separate relaxations introduced by Wang et al. (Wang et al., 2018b) and combine them with the adaptive ReLU relaxations published in (Singh et al., 2019b; Zhang et al., 2018) and propagate the symbolic bounds through

$$\overline{\text{ReLU}}(x) = \frac{u_u}{u_u - u_l} x - \frac{u_u u_l}{u_u - u_l} \quad (4)$$

$$\underline{\text{ReLU}}(x) = \begin{cases} 0, & l_l < |l_l| \\ x, & l_l \geq |l_l| \end{cases}, \quad (5)$$

where $[l_l, l_u]$ and $[u_l, u_u]$ are concrete intervals enclosing the values of $lb(\mathbf{x})$ and $ub(\mathbf{x})$ respectively.

If the satisfiability or unsatisfiability of a property cannot be established from a single symbolic forward pass, we make use of the branch and bound implementation developed by Strong et al. (Strong et al., 2021) based on strong branching and splitting on the input domain. Although the result is not a complete verifier for ReLU networks, it rapidly reduces the obtained bounds for low-dimensional input spaces. In order to increase the scalability of the branch and bound approach, we devise a new splitting heuristic targeted at reducing overapproximation of unstable neurons, presented in Sec. 3.2. In Sec. 3.3, we present two further optimizations. The first one ensuring monotonically tighter bounds

for smaller sub-problems, the second one increasing the efficiency of finding counterexamples or good candidate solutions for optimization problems.

3.1. New Heuristics for Introducing Fresh Variables

Paulsen et al. (Paulsen et al., 2020) showed with their tool NEURODIFF that introducing fresh variables can significantly improve performance. But they also note that introducing a fresh variable removes all occurrences of previous variables in the symbolic bounds with just the fresh variable. Therefore introducing a fresh variable diminishes the chances of earlier variables to cancel. Furthermore, each fresh variable increases the size of the symbolic bounds and consequently makes propagation through the network more expensive. While we decide to treat the maximum total number of fresh variables as a hyperparameter, Paulsen et al. (Paulsen et al., 2020) limit the number of fresh variables to $n_{\text{var}} = \sum_{l=1}^L \frac{1}{l} N_l$, where N_l is the number of unstable neurons in layer l . To mitigate the first problem in particular, they introduce these n_{var} variables for the earliest possible unstable neurons in the network.

However, this heuristic does not perform well, when a whole layer of the network consists of only unstable neurons (or fixed-zero and unstable neurons). In this extreme case all references to earlier variables are removed from the symbolic bounds. Motivated by this example, we introduce a parameter $\lambda \in [0, 1]$ to bound the number of fresh variables introduced in layer l by

$$n_{\text{var}_l} = \lambda \cdot n_l, \quad (6)$$

where n_l is the number of neurons in that layer. If λ is low, we introduce fewer variables per layer, but existing variables in earlier layers (including the symbolic input variables) have a higher probability of cancelling out later

on. The effect of this optimization, when we perform a symbolic forward pass for the ACAS-Xu.1.1 network with increasing maximal number of fresh variables and different values of λ , can be seen on the left of Fig. 2. The heuristic of (Paulsen et al., 2020) ($\lambda = 1$) initially improves upon the upper bound obtained from symbolic forward propagation without additional variables, but when too many variables are added, performance becomes significantly worse, as the new variables interfere with each others' cancellation. For smaller values of λ , there is no such error-explosion effect, even when a large number of fresh variables is introduced.

While the parameter λ controls the maximal fraction of newly introduced variables in a layer, it does not determine *which* neurons on a layer are selected. A further improvement of our heuristic thus selects the neurons on each level, for which new variables are introduced.

The idea is as follows: The main advantage of introducing fresh variables is that they have the opportunity to cancel out in later stages of the symbolic forward pass, such that either their symbolic lower or their symbolic upper bound – never (implicitly) both – are used in the calculation of concrete bounds for later neurons. Implicitly using both symbolic bounds of a neuron whose symbolic bounds are far apart likely creates worse bounds than using both symbolic bounds of a neuron whose symbolic lower and upper bound only minimally differ. Approximating this behaviour, we prioritize introduction of fresh variables for unstable neurons with large ranges between *concrete* lower and upper bounds.

Additionally, we introduce two minor optimizations. First, instead of directly using the number of neurons in layer l as n_l in Equation 6, we choose n_l to be the number of fixed-active or unstable neurons, excluding fixed-inactive neurons, as their symbolic bounds are all zero. As our second improvement, we prohibit the introduction of fresh variables in the last hidden layer, as these variables have no chance to cancel out before the output layer. In practice, these two improvements have only a minor effect, as there are few layers with no fixed-active neurons, and when a maximum number of variables is set, there are rarely ones left to be introduced in the last hidden layer.

The overall effect of our fresh-variable-selection-heuristic and the two small optimizations can be seen on the right of Fig. 2, where the solid lines represent the upper bounds obtained from a single symbolic forward pass for property ϕ_1 through the ACAS-Xu.1.1 network for some parameters λ also shown on the left of the figure.

3.2. Splitting Heuristic

Introducing fresh variables not only results in tighter bounds obtained from single symbolic forward passes. By obtaining

better bounds, it can also reduce the number of passes required by a branch and bound algorithm to prove a property, often dramatically, thus compensating the slightly higher computation time.

Performance of the branch and bound algorithm can be further improved by better splitting decisions. Existing input splitting methods have focused on splitting the input with the largest range between its lower and upper bound (Strong et al., 2021) while others have used gradient information to bisect the input x_i with the largest smear value (Wang et al., 2018a). Although the gradient captures information about the importance of an input variable on the output of the network, neither the gradient, nor the width of the input interval are a good estimate for the amount of overapproximation-error an input variable is responsible for. More recently, Henriksen and Lomuscio (Henriksen & Lomuscio, 2021) reported significant improvements in verification times, utilizing the error terms in error-based symbolic interval propagation to take interval overapproximation into account. However, their heuristic is restricted to parallel relaxations for the activation functions.

We therefore adapt their approach and make use of the direct effect of the input variables on the bounds of intermediate neurons. Given inputs \mathbf{x} with $x_i \in [\underline{x}_i, \bar{x}_i]$ and a neuron n_j with symbolic lower and upper bounds $L_j(\mathbf{x}) = l_{0j} + \sum_i l_{ij}x_i$ and $U_j(\mathbf{x}) = u_{0j} + \sum_i u_{ij}x_i$ and concrete bounds $[\underline{L}_j, \bar{U}_j]$, we can compute the effect of bisecting the range of input x_i into two intervals $[\underline{x}_i, \hat{x}_i]$ and $[\hat{x}_i, \bar{x}_i]$, where $\hat{x}_i = 1/2(\underline{x}_i + \bar{x}_i)$, on the concretization of $[L_j(\mathbf{x}), U_j(\mathbf{x})]$. After bisection, we obtain new concrete bounds $\underline{L}'_j(\mathbf{x})$ and $\bar{U}'_j(\mathbf{x})$:

$$\underline{L}'_j(\mathbf{x}) = \begin{cases} \underline{L}_j(\mathbf{x}), & l_{ij} \geq 0, x_i \in [\underline{x}_i, \hat{x}_i] \\ \underline{L}_j(\mathbf{x}) - \frac{l_{ij}}{2}(\bar{x}_i - \underline{x}_i), & l_{ij} \leq 0, x_i \in [\underline{x}_i, \hat{x}_i] \\ \underline{L}_j(\mathbf{x}) + \frac{l_{ij}}{2}(\bar{x}_i - \underline{x}_i), & l_{ij} \geq 0, x_i \in [\hat{x}_i, \bar{x}_i] \\ \underline{L}_j(\mathbf{x}), & l_{ij} \leq 0, x_i \in [\hat{x}_i, \bar{x}_i] \end{cases}$$

$$\bar{U}'_j(\mathbf{x}) = \begin{cases} \bar{U}_j(\mathbf{x}) - \frac{u_{ij}}{2}(\bar{x}_i - \underline{x}_i), & u_{ij} \geq 0, x_i \in [\underline{x}_i, \hat{x}_i] \\ \bar{U}_j(\mathbf{x}), & u_{ij} \leq 0, x_i \in [\underline{x}_i, \hat{x}_i] \\ \bar{U}_j(\mathbf{x}), & u_{ij} \geq 0, x_i \in [\hat{x}_i, \bar{x}_i] \\ \bar{U}_j(\mathbf{x}) + \frac{u_{ij}}{2}(\bar{x}_i - \underline{x}_i), & u_{ij} \leq 0, x_i \in [\hat{x}_i, \bar{x}_i] \end{cases}$$

In each of these cases, there is at least one branch where at least one of the bounds is improved and the magnitude of the improvement depends on the range $\bar{x}_i - \underline{x}_i$ of the input x_i as well as the *magnitude* $|u_{ij}|$ and $|l_{ij}|$ of the coefficients of x_i in the symbolic lower and upper bound.

In order to reduce the internal overapproximation, we want to split on inputs that cause a large improvement of the bounds of many intermediate neurons. The importance of

an input x_i is therefore determined as

$$s(x_i) = \sum_{j=1}^k \frac{1}{2} (\bar{x}_i - \underline{x}_i) (|l_{ij}| + |u_{ij}|), \quad (7)$$

where the n_j are the k unstable neurons in the network. The heuristic then selects the input with largest $s(x_i)$ for splitting.

3.3. Additional Optimizations

If we bisect the input of a NN verification problem to obtain two smaller sub-problems, it is immediately clear that the bounds obtained for the sub-problems have to be at least as tight as the bounds obtained on the original problem, as we maximize or minimize over a smaller input set. Our symbolic interval propagation algorithm, however, does not naturally return tighter bounds for the sub-problems, as the adaptive ReLU relaxations proposed by (Singh et al., 2019b; Zhang et al., 2018) and the heuristics for introducing fresh variables are sources of non-continuity.

If we split on an input, and thus the concrete bounds on some ReLU neuron change, the adaptive lower relaxation might change accordingly and while it still minimizes the area of overapproximation at this particular neuron, it might negatively affect the tightness of the bounds of later neurons.

Similarly for our fresh variable introduction heuristic: If the concrete bounds of the neurons change due to smaller input spaces, we might introduce fresh variables for different neurons in the sub-problem than in the original problem. The variables for these neurons might cancel less favourably and thus lead to looser output bounds.

In order to eliminate this behaviour of not monotonically improving bounds, we can store the output bounds of the original problem and only update the bounds in the sub-problem, if those bounds are actually tighter than before. However, since we minimize or maximize over a smaller subset of the input space not only for the final outputs of the NN, but also for all of its interior neurons, we can proceed in the same manner for these interior neurons and also only update their bounds if they are improved as well.

While the previous optimizations were all concerned with improving the *upper bound* on the maximum value of the optimization problem 3, we also propose a heuristic to obtain good candidates for maximizing inputs or counterexamples \mathbf{x} , whose value $f_{\mathcal{N}}(\mathbf{x})$ serves as a *lower bound* to the maximal attainable value of the NN. To generate candidates for \mathbf{x} , ZOPE (Strong et al., 2021) periodically samples the center point of the input hyperrectangle of the current sub-problem. NEURIFY (Wang et al., 2018b) on the other hand, uses the input vector that maximizes a linear program containing the symbolic output bounds of the NN and constraints fixing the state of certain intermediate ReLU neurons. The main idea

behind this, is that the resulting linear program is a good approximation of the underlying NN, if there is low overapproximation. In input-splitting however, no additional linear constraints arise from fixing the state of ReLU neurons. Therefore the linear program reduces to just the symbolic interval enclosing the NN’s outputs. We thus make the same assumption that these symbolic intervals $[lb(\mathbf{x}), ub(\mathbf{x})]$ are a good approximation of the true outputs of the NN, and use the input that maximizes the symbolic upper bound $ub(\mathbf{x})$ as candidate for a maximizing input. This operation comes without any additional computational cost, as the calculation of the concrete upper bound \bar{y} of the output $f_{\mathcal{N}}(\mathbf{x})$ of the NN already depends on setting the corresponding input variables to their appropriate value.

4. Evaluation

The implementation in Julia of our approach¹ is called DPNEURIFYFV (for NEURIFY (Wang et al., 2018b) with fresh variables and the ReLU relaxations used in DEEPPOLY (Singh et al., 2019b)). It builds upon `NeuralVerification.jl`² (Liu et al., 2021) and ZOPE³ (Strong et al., 2021).

In our evaluation, we want to investigate, how DPNEURIFYFV compares to state of the art tools in NN verification. Additionally, we want to gain further insights into the effect of our proposed optimizations. All experiments were run on a single thread and were conducted on a desktop computer running a 64 bit version of Ubuntu 20.04.1 with an Intel Core i7-2600 3.4 GHz processor, which has 8 cores, and 8GB of RAM.

4.1. The ACAS Xu Benchmark.

To demonstrate the effectiveness of our approach, we use the ACAS Xu neural network verification benchmark set published by Katz et al. (Katz et al., 2017). The benchmark consists of NNs that were trained to provide navigation advisories on board of aeroplanes as part of the Airborne Collision Avoidance System for unmanned aircraft. Each of the 45 NNs has 5 input neurons taking in parameters of the own aircraft and aircraft nearby, 6 fully connected layers of 50 neurons each, and 5 output neurons representing scores for the different navigation advisories. We consider properties ϕ_1, ϕ_2, ϕ_3 and ϕ_4 of the benchmark set, which are to be checked for each of the 45 networks. All of the properties define a hyperrectangle of valid values for the input space and an appropriate property on the output space of the NN.

¹Code available at <https://github.com/phK3/DPNeurifyFV.jl>

²Code available at <https://github.com/sisl/NeuralVerification.jl>.

³Code available at <https://github.com/sisl/NeuralPriorityOptimizer.jl>.

Property ϕ_1 asks for an upper bound on the first output of the network and can therefore be modelled as maximizing the violation of a single linear constraint, while properties ϕ_2, ϕ_3 and ϕ_4 require that certain outputs never have the minimal or maximal score among all outputs of the network. As the constraints $o_i \geq o_j, \forall i \neq j$, where the o_k are the outputs of the NN, represent a convex polytope, we want to ensure that the output set of the NN never intersects this polytope. Therefore, these latter properties can be modelled by minimizing the violation of the associated constraints of the polytope.

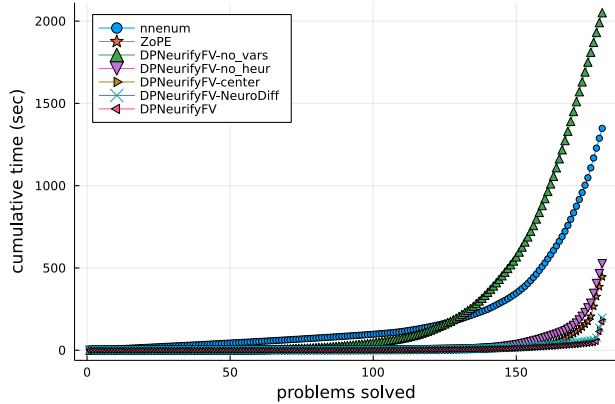


Figure 3: Overall solving time for properties $\phi_1, \phi_2, \phi_3, \phi_4$ over all networks of the ACAS-Xu family for ZOPE, NNENUM, our tool DPNEURIFYFV and ablations of our tool without introducing fresh variables, without utilizing our splitting heuristic, sampling the center point instead of the maximizer/minimizer of the symbolic bound to obtain a candidate solution and using the NEURODIFF heuristic for fresh variable introduction.

4.2. Verification.

We run our tool DPNEURIFYFV on all the aforementioned ACAS Xu verification tasks, and compare its performance to ZOPE (Strong et al., 2021), a solver for optimization problems over the output of NNs and NNENUM⁴ (Bak et al., 2020), the NN verifier that scored highest amongst all participants at VNN-COMP 2021 (Bak et al., 2021) on the ACAS Xu benchmark set. We configured DPNEURIFYFV to introduce at most 20 fresh variables and to introduce fresh variables for a maximal fraction $\lambda = 1/2$ of a layer’s non-fixed zero neurons. For ZOPE, we used the setting the authors published in the appendix of their paper (Strong et al., 2021), while we used the standard parameters for the configuration of NNENUM (Bak et al., 2020), only changing the number of processes to one. For all methods, we set a timeout of 60 seconds per instance. The overall

⁴Code available at <https://github.com/stanleybak/nnenum>.

result is shown in Table 1 and displayed as a cactus plot in Fig. 3. The results indicate that DPNEURIFYFV is significantly faster than NNENUM. While NNENUM finishes verification of 175 of the 180 tasks in 1,048.2 seconds, our tool successfully verifies 178 of the properties in just 52.73 seconds – a $19\times$ improvement. In comparison to ZOPE’s runtime of 205.9 seconds for the verification of 176 properties, our approach is still considerably faster by a factor of 4. when minimizing the violation of the associated constraints.

To investigate the effect of our different optimizations, we compare the full implementation of our approach to variants without certain optimizations. We refer to the version of DPNEURIFYFV without the introduction of fresh variables as DPNEURIFYFV-NO_VARS, without the usage of our splitting heuristic as DPNEURIFYFV-NO_HEUR, to the version without using the maximizer/minimizer of the symbolic output bounds to find good solution candidates as DPNEURIFYFV-CENTER, as it just samples the center point of the input hyperrectangle of the current sub-problem and to the version that uses the heuristic of NEURODIFF (Paulsen et al., 2020) for introduction of fresh variables as DPNEURIFYFV-NEURODIFF.

The effect of introducing fresh variables is the largest improvement by far, as can be seen from the performance of DPNEURIFYFV-NO_VARS, which is significantly slower than NNENUM on both satisfiable as well as unsatisfiable instances. Due to the dependency issue, the obtained symbolic bounds do neither tightly approximate the true optimization target, nor the input bounds to the intermediate neurons, thus leading to large overapproximations.

5. Conclusion and Future Work

In this paper, we presented DPNEURIFYFV, an algorithm for verification of ReLU NNs with low-dimensional input space. The approach is based on input splitting and combining several methods for symbolic interval propagation. We added important further optimizations encompassing new heuristics for introducing fresh variables, a novel splitting heuristic that can be used for a wide set of ReLU relaxations and further smaller improvements of the branch-and-bound search.

We evaluated DPNEURIFYFV on the ACAS Xu benchmark set and compared it to the state-of-the-art verification tools ZOPE and NNENUM, achieving considerable improvements in runtime. We also empirically evaluated the effect of our different improvements, showing that introducing new variables is essential, but often only in combination with further optimizations. We also empirically evaluated the effect of our different improvements, showing that introducing new variables is essential, but often only in combination with other optimizations.

Table 1: Number of solved instances and cumulative verification times for properties ϕ_1, ϕ_2, ϕ_3 and ϕ_4 of the ACAS Xu verification benchmark (180 instances total).

Approach	SAT	(sec)	UNSAT	(sec)	INCONCLUSIVE
NNENUM (Bak et al., 2020)	44	114.09	131	934.11	5
ZOPE (Strong et al., 2021)	44	81.51	132	124.39	4
DPNEURIFYFV-no_vars	45	137.33	124	1,253.53	11
DPNEURIFYFV-no_heur	45	100.17	132	245.58	3
DPNEURIFYFV-center	45	25.78	133	40.48	2
DPNEURIFYFV-NEURODIFF	45	20.40	133	56.02	2
DPNEURIFYFV	45	12.82	133	39.91	2

Future work might include an adaptation of our method to the equivalence verification problem for NNs (Kleine Büning et al., 2020; Teuber et al., 2021) as well as an evaluation of our approach on further benchmarks.

Acknowledgements

The authors would like to thank the Ministry of Science, Research and Arts of the Federal State of Baden-Württemberg, Germany for financial support within the InnovationCampus Future Mobility.

References

- Bak, S., Tran, H., Hobbs, K., and Johnson, T. T. Improved geometric path enumeration for verifying ReLU neural networks. In *CAV (I)*, volume 12224 of *Lecture Notes in Computer Science*, pp. 66–96. Springer, 2020.
- Bak, S., Liu, C., and Johnson, T. The second international verification of neural networks competition (VNN-COMP 2021): Summary and results, 2021.
- Bojarski, M., Testa, D. D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L. D., Monfort, M., Muller, U., Zhang, J., Zhang, X., Zhao, J., and Zieba, K. End to end learning for self-driving cars, 2016.
- Bunel, R., Lu, J., Turkaslan, I., Torr, P. H. S., Kohli, P., and Kumar, M. P. Branch and bound for piecewise linear neural network verification. *J. Mach. Learn. Res.*, 21: 42:1–42:39, 2020.
- Cheng, C.-H., Nührenberg, G., and Ruess, H. Maximum resilience of artificial neural networks. *Lecture Notes in Computer Science*, pp. 251–268, 2017. ISSN 1611-3349. doi: 10.1007/978-3-319-68167-2_18. URL http://dx.doi.org/10.1007/978-3-319-68167-2_18.
- Ehlers, R. Formal verification of piece-wise linear feed-forward neural networks. In D’Souza, D. and Narayan Kumar, K. (eds.), *Automated Technology for Verification and Analysis*, pp. 269–286, Cham, 2017. Springer International Publishing. ISBN 978-3-319-68167-2.
- Fischer, M., Balunovic, M., Drachler-Cohen, D., Gehr, T., Zhang, C., and Vechev, M. D12: Training and querying neural networks with logic. In *International Conference on Machine Learning*, pp. 1931–1941, 2019.
- Gehr, T., Mirman, M., Drachler-Cohen, D., Tsankov, P., Chaudhuri, S., and Vechev, M. Ai2: Safety and robustness certification of neural networks with abstract interpretation. In *2018 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18. IEEE, 2018.
- Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>.
- Henriksen, P. and Lomuscio, A. DEEPSPLIT: an efficient splitting method for neural network verification via indirect effect analysis. In Zhou, Z. (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pp. 2549–2555. ijcai.org, 2021. doi: 10.24963/ijcai.2021/351. URL <https://doi.org/10.24963/ijcai.2021/351>.
- Henriksen, P. and Lomuscio, A. R. Efficient neural network verification via adaptive refinement and adversarial search. In *ECAI*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pp. 2513–2520. IOS Press, 2020.
- Huang, X., Kwiatkowska, M., Wang, S., and Wu, M. Safety verification of deep neural networks. In *International conference on computer aided verification*, pp. 3–29. Springer, 2017.
- Julian, K. D., Lopez, J., Brush, J. S., Owen, M. P., and Kochenderfer, M. J. Policy compression for aircraft collision avoidance systems. In *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)*, pp. 1–10. IEEE, 2016.

- Katz, G., Barrett, C., Dill, D. L., Julian, K., and Kochenderfer, M. J. Reluplex: An efficient smt solver for verifying deep neural networks. *Lecture Notes in Computer Science*, pp. 97–117, 2017. ISSN 1611-3349. doi: 10.1007/978-3-319-63387-9_5. URL http://dx.doi.org/10.1007/978-3-319-63387-9_5.
- 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, 2019.
- Kleine Büning, M., Kern, P., and Sinz, C. Verifying equivalence properties of neural networks with ReLU activation functions. In *Principles and Practice of Constraint Programming*, pp. 868–884. Springer, 2020. ISBN 978-3-030-58475-7.
- Liu, C., Arnon, T., Lazarus, C., Strong, C., Barrett, C., and Kochenderfer, M. J. Algorithms for verifying deep neural networks. *Foundations and Trends® in Optimization*, 4(3-4):244–404, 2021. ISSN 2167-3888. doi: 10.1561/24000000035. URL <http://dx.doi.org/10.1561/24000000035>.
- Moore, R. E. *Interval analysis*, volume 4. Prentice-Hall Englewood Cliffs, 1966.
- Müller, C., Serre, F., Singh, G., Püschel, M., and Vechev, M. Scaling polyhedral neural network verification on gpus. *Proceedings of Machine Learning and Systems*, 3: 733–746, 2021a.
- Müller, M. N., Makarchuk, G., Singh, G., Püschel, M., and Vechev, M. Prima: Precise and general neural network certification via multi-neuron convex relaxations. *arXiv preprint arXiv:2103.03638*, 2021b.
- Paulsen, B., Wang, J., Wang, J., and Wang, C. NEURO-DIFF: scalable differential verification of neural networks using fine-grained approximation. In *35th IEEE/ACM International Conference on Automated Software Engineering, ASE 2020, Melbourne, Australia, September 21-25, 2020*, pp. 784–796. IEEE, 2020. doi: 10.1145/3324884.3416560. URL <https://doi.org/10.1145/3324884.3416560>.
- Singh, G., Gehr, T., Mirman, M., Püschel, M., and Vechev, M. T. Fast and effective robustness certification. In *NeurIPS*, pp. 10825–10836, 2018.
- Singh, G., Ganvir, R., Püschel, M., and Vechev, M. Beyond the single neuron convex barrier for neural network certification. *Advances in Neural Information Processing Systems*, 32, 2019a.
- Singh, G., Gehr, T., Püschel, M., and Vechev, M. T. An abstract domain for certifying neural networks. *Proc. ACM Program. Lang.*, 3(POPL):41:1–41:30, 2019b.
- Stolfi, J. and Figueiredo, L. H. D. Self-validated numerical methods and applications, 1997.
- Strong, C. A., Katz, S. M., Corso, A. L., and Kochenderfer, M. J. Zope: A fast optimizer for ReLU networks with low-dimensional inputs. *CoRR*, abs/2106.05325, 2021. URL <https://arxiv.org/abs/2106.05325>.
- Teuber, S., Kleine Büning, M., Kern, P., and Sinz, C. Geometric path enumeration for equivalence verification of neural networks. In *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 200–208, 2021. doi: 10.1109/ICTAI52525.2021.00035.
- Tjeng, V., Xiao, K., and Tedrake, R. Evaluating robustness of neural networks with mixed integer programming, 2017.
- Tran, H.-D., Yang, X., Manzananas Lopez, D., Musau, P., Nguyen, L. V., Xiang, W., Bak, S., and Johnson, T. T. Nnv: the neural network verification tool for deep neural networks and learning-enabled cyber-physical systems. In *International Conference on Computer Aided Verification*, pp. 3–17. Springer, 2020.
- Wang, S., Pei, K., Whitehouse, J., Yang, J., and Jana, S. Formal security analysis of neural networks using symbolic intervals. In *USENIX Security Symposium*, pp. 1599–1614. USENIX Association, 2018a.
- Wang, S., Pei, K., Whitehouse, J., Yang, J., and Jana, S. Efficient formal safety analysis of neural networks. In *NeurIPS*, pp. 6369–6379, 2018b.
- Xiang, W., Tran, H.-D., and Johnson, T. T. Output reachable set estimation and verification for multilayer neural networks. *IEEE transactions on neural networks and learning systems*, 29(11):5777–5783, 2018.
- Zhang, H., Weng, T., Chen, P., Hsieh, C., and Daniel, L. Efficient neural network robustness certification with general activation functions. In *NeurIPS*, pp. 4944–4953, 2018.