Speeding up neural network robustness verification via algorithm configuration and an optimised mixed integer linear programming solver portfolio

König, H.M.T.; Hoos, H.H.; Rijn, J.N. van

Citation

König, H. M. T., Hoos, H. H., & Rijn, J. N. van. (2022). Speeding up neural network robustness verification via algorithm configuration and an optimised mixed integer linear programming solver portfolio. *Machine Learning*. doi:10.1007/s10994-022-06212-w

Version: Publisher's Version

License: Creative Commons CC BY 4.0 license

Downloaded from: https://hdl.handle.net/1887/3484755

Note: To cite this publication please use the final published version (if applicable).

Machine Learning
https://doi.org/10.1007/s10994-022-06212-w

Speeding up neural network robustness verification via algorithm configuration and an optimised mixed integer linear programming solver portfolio

Matthias König[1] · Holger H. Hoos[1,2] · Jan N. van Rijn[1]

Received: 14 October 2021 / Revised: 11 April 2022 / Accepted: 8 June 2022
© The Author(s) 2022

Abstract
Despite their great success in recent years, neural networks have been found to be vulnerable to adversarial attacks. These attacks are often based on slight perturbations of given inputs that cause them to be misclassified. Several methods have been proposed to formally prove robustness of a given network against such attacks. However, these methods typically give rise to high computational demands, which severely limit their scalability. Recent state-of-the-art approaches state the verification task as a minimisation problem, which is formulated and solved as a mixed-integer linear programming (MIP) problem. We extend this approach by leveraging automated algorithm configuration techniques and, more specifically, construct a portfolio of MIP solver configurations optimised for the neural network verification task. We test this approach on two recent, state-of-the-art MIP-based verification engines, MIPVerify and Venus, and achieve substantial improvements in CPU time by average factors of up to 4.7 and 10.3, respectively.

Keywords Neural network verification · Mixed integer programming · Automated algorithm configuration · Algorithm selection

Editors: Krzysztof Dembczynski and Emilie Devijver.

✉ Matthias König
h.m.t.konig@liacs.leidenuniv.nl

Holger H. Hoos
hh@liacs.nl

Jan N. van Rijn
j.n.van.rijn@liacs.leidenuniv.nl

[1] Leiden Institute of Advanced Computer Science, Leiden University, Leiden, The Netherlands

[2] University of British Columbia, Vancouver, Canada

Published online: 08 September 2022

1 Introduction

In recent years, deep learning algorithms have become increasingly important tools within various application domains and use contexts, ranging from object recognition systems in autonomous cars to face recognition systems in mobile phones. At the same time, it is now well known that neural networks are vulnerable to *adversarial attacks* (Szegedy et al., 2014), in which a given input is transformed in such a way that it is misclassified by the network. In the case of image recognition tasks, the required perturbation can be so small that it remains virtually undetectable to the human eye.

Various methods have been proposed to establish robustness of neural networks against adversarial attacks. Some of these methods perform *heuristic* attacks (Goodfellow et al., 2014; Kurakin et al., 2016; Carlini & Wagner, 2017); however, these do not paint a full picture of a given network's robustness to adversarial attacks, as one defense mechanism might still be circumvented by another, possibly new class of attacks. In light of this, approaches have been developed to more thoroughly verify neural networks (Scheibler et al., 2015; Bastani et al., 2016; Ehlers, 2017; Katz et al., 2017; Dvijotham et al., 2018; Gehr et al., 2018; Xiang et al., 2018; Bunel et al., 2018; Tjeng et al., 2019; Botoeva et al., 2020). These *formal* verification methods can assess the robustness of a given network in a principled fashion, which means that they yield provable guarantees on certain properties of input-output combinations. However, this class of network verification methods tends to be computationally expensive, making it difficult to verify networks with a large number of units and/or on a large number of inputs.

Recent work by Tjeng et al. (2019) addressed this challenge and presented a verification tool, called MIPVerify, that, for the first time, was able to evaluate the robustness of larger neural networks on the full MNIST dataset. In their study, Tjeng et al. (2019) formulate the verification task as a minimisation problem, which is then solved through *mixed-integer linear programming (MIP)*. More specifically, the optimisation task is to apply a perturbation to the original sample that maximises model error, while staying close to the initial example, i.e., keeping the distance at a minimum. In other words, the verifier takes an image and a trained neural network as inputs and produces either an adversarial example or, if the optimisation problem cannot be solved, a certificate of local robustness. While MIPVerify can verify a larger number of instances than previous methods, such as those from the works of Wong and Kolter (2018), Dvijotham et al. (2018) or Raghunathan et al. (2018), it is computationally costly (in terms of CPU time required per verification query). Specifically, depending on the classifier to be verified, we found that some instances required several thousand CPU seconds of running time of the MIP solver, while a sizeable fraction of instances could not be solved at all, even within a rather generous time limit of 38 400 CPU seconds per sample.

The same holds for other MIP-based verification systems, such as Venus by Botoeva et al. (2020), which has been demonstrated to be faster than many other state-of-the-art verification tools, including the MIP-based verifier NSVerify Akintunde et al. (2018). Here, our experiments showed that, depending on the classifier to be verified, the computational cost per query remains subject to great variance as outlined above, with many instances resulting in timeouts.

We note that, to date, the performance of MIPVerify and Venus has not been compared directly, which motivates our decision to consider both as contributors to the state of the art in MIP-based neural network verification.

Previous work has demonstrated that automated configuration of MIP solvers can yield substantial improvements (Hutter et al., 2009, 2010, 2011; Lopez-Ibanez & Stützle, 2014). Building on these findings, we seek to improve the performance of MIP-based neural network verification tools by leveraging *automated algorithm configuration* techniques to optimise the hyperparameters of the solver at the heart of these verifiers. As such, the proposed method can be used regardless of the underlying MIP problem formulation, and its improvements are orthogonal to any advances made in this regard. Put differently, we argue that automated algorithm configuration can benefit any verification approach relying on MIP solving or similar techniques.

Automated algorithm configuration of neural network verification engines is a non-trivial task and comes with its own challenges. Most prominently, the high running times and heterogeneity/diversity of instances pose problems that are not easily solved by standard configuration approaches, such as SMAC (Hutter et al., 2011). More precisely, we consistently found in our experiments that a single configuration could not significantly improve mean CPU time over the default. In fact, we observed that a single configuration could achieve a 500-fold speedup on a given instance over the default, but then time out on another, which the default, in turn, could solve.

Therefore, we decided to adapt Hydra (Xu et al., 2010), an advanced approach that combines algorithm configuration and per-instance algorithm selection, to automatically construct a *parallel portfolio* of MIP solver configurations optimised for solving neural network verification problems.

We demonstrate the effectiveness of our approach for both aforementioned verification tools. These systems both rely on MIP solving, yet they are conceptually different enough to show the generalisability of our method. This study can be seen as an extension of our recent workshop publication on the same topic (König et al., 2021), in which we reported preliminary results for MIPVerify on a single benchmark. To the best of our knowledge, ours is the first study to pursue this direction. In brief, the main contributions are as follows:

- A framework for automatically constructing a parallel portfolio of MIP solver configurations optimised for neural network verification, which can be applied to any MIP-based verification method.
- An extensive evaluation of this framework on two the state-of-the-art verification engines, namely Venus (Botoeva et al., 2020) and MIPVerify, improving their performance on (i) SDP_dMLP_A - an MNIST classifier designed for robustness (Raghunathan et al., 2018), (ii) mnistnet - an MNIST classifier from the neural network verification literature (Botoeva et al., 2020) and (iii) the ACAS Xu benchmark (Julian et al., 2016; Katz et al., 2017).

On the SDP_dMLP_A benchmark, we achieve substantial improvements in CPU time by average factors of 4.7 and 10.3 for MIPVerify and Venus, respectively, over the state of the art on a *solvable* subset of instances from the MNIST dataset. This subset excludes all instances that cannot be solved by any of the baseline approaches we consider. Beyond that, the number of timeouts was reduced by a factor of 1.42 and 1.6, respectively.

On the mnistnet benchmark, we again achieved substantial improvements in CPU time, this time by average factors of 1.61 and 7.26 for MIPVerify and Venus, respectively, over the state of the art on solvable instances. We furthermore reduced timeouts on this benchmark by average factors of 1.14 and 2.81, respectively.

Finally, we strongly improved the performance of the Venus verifier on the ACAS Xu benchmark, attaining a 2.97-fold reduction in average CPU time. We note that on this benchmark, we found MIPVerify to be unable to solve most of the instances within the kinds of computational budgets considered in our experiments.

2 Background

The following section provides an overview of adversarial examples and methods to verify neural networks against them. It further puts focus on the limitations of current state-of-the-art approaches and introduces the concepts behind automated algorithm configuration and portfolio construction.

2.1 Adversarial examples

Adversarial examples or negatives are network inputs that are indistinguishable from regular inputs, but cause the network to produce misclassifications (Szegedy et al., 2014). These adversarial examples can be produced by applying a hardly perceptible perturbation to the original instance that maximises model error while staying close to the initial example. The most prevalent distance metrics used to evaluate adversarial distortion are l_1 (Carlini et al,. 2017; Chen et al. 2018), l_2 (Szegedy et al., .,2014) and l_∞ (Goodfellow et al., 2014; Papernot et al., 2016) norm.

2.2 Network verification

Several methods have been produced to evaluate neural networks through heuristic attacks (Goodfellow et al., 2014; Kurakin et al., 2016; Carlini & Wagner, 2017). However, these algorithms cannot accurately assess network robustness. That is, a classifier trained to be robust against one class of attacks can still be vulnerable to another (Carlini et al., 2017).

To tackle this problem, more advanced techniques have been introduced for the *formal verification* of neural networks (Scheibler et al., 2015; Bastani et al., 2016; Ehlers, 2017; Katz et al., 2017; Dvijotham et al., 2018; Gehr et al., 2018; Xiang et al., 2018; Bunel et al., 2018; Tjeng et al., 2019; Botoeva et al., 2020). These methods verify whether a particular network satisfies certain input-output properties or provide an example for which the property is violated. For a classifier, a property can be that instances, which are in close distance to a certain input x, belong to the same class as x.

In general, formal verification algorithms can be characterised by three criteria: *soundness, completeness* and computational cost. A sound algorithm will only report that a property holds if the property actually holds. An algorithm that is complete will correctly state that a property holds whenever it holds. While it is favourable to produce verifiers that can certify every given instance in a dataset, there is a trade-off between completeness of a verification algorithm and its scalability in terms of computational complexity. Neural network verification is highly complex, and even simple properties about them have been proven to be NP-complete problems (Katz et al., 2017). This makes it intractable to apply complete verification techniques to large networks and/or instance sets.

Consequently, some verification algorithms forego completeness to improve computational efficiency by making approximations (Bastani et al., 2016; Dvijotham et al., 2018; Gehr et al., 2018; Xiang et al., 2018; Bunel et al., 2018). These approximations, however,

do not always return the actual solution to a given verification problem but can result in mismatches or cases, where the solution remains unknown. Other incomplete methods seek to add random noise to smooth a neural network classifier and then derive the certified robustness for this smoothed classifier (Lecuyer et al., 2019; Cohen et al., 2019). While these approaches scale to larger network architectures, their robustness guarantees remain probabilistic. Furthermore, randomised smoothing has been found to come at the cost of classifier accuracy (Mohapatra et al., 2021). As can be seen from this example, increased scalability of a verification method usually comes at the cost of performance loss in other areas.

Recent work by Tjeng et al. (2019) seeks to overcome this trade-off by presenting a verifier that is complete and scalable to larger neural networks. Their verifier, MIPVerify, combines and extends existing approaches to MIP-based robustness verification (Cheng et al., 2017; Lomuscio & Maganti, 2017; Dutta et al., 2018; Fischetti & Jo, 2018) and presents a verifier that encodes the network as a set of mixed-integer linear constraints. Following Tjeng et al. (2019), a valid adversarial example x' for input x with true class label $\lambda(x)$ (encoded as integer) corresponds to the solution to the problem where we minimize:

$$d(x', x) \tag{1}$$

subject to

$$\arg\max_i(f_i(x')) \neq \lambda(x) \tag{2}$$

$$x' \in (G(x) \cap X_{valid}) \tag{3}$$

where $d(\cdot,\cdot)$ denotes a distance metric (e.g., the l_∞-norm), $f_i(\cdot)$ is the i-th network output (i.e., indicating whether it predicts the input to belong to the i-th class) and $G(x) = \{x' \mid \forall i : -\varepsilon \leq (x - x')_i \leq \varepsilon\}$. Intuitively, $G(x)$ denotes the region around an input x corresponding to all allowable perturbations within a pre-defined radius ε. X_{valid} represents the domain of valid inputs (e.g., the pixel value range of a normalised image, in case of image classification). Note that this formulation assumes that the network predicts a single class label for each observation (i.e., the arg max operator in Eq. 2 returns a single element); other behaviour is undefined.

MIPVerify achieves speed-ups through optimised MIP formulations or, more specifically, tight formulations for non-linearities and a pre-solving algorithm that reduces the number of binary variables, i.e., the number of unstable ReLU nodes. More specifically, the information provided by $G(x)$ is used to reduce the interval of the input domain propagated through the network during the calculation of the pre-activation bounds. This is combined with *progressive bounds tightening*, which represents a method for choosing procedures to determine pre-activation bounds, i.e., interval arithmetic or linear programming, based on the potential improvement to the problem formulation.

The MIP-based verifier Venus (Botoeva et al., 2020) achieves performance gains over previous methods, such as NSVerify (Akintunde et al., 2018), through dependency-based pruning to reduce the search space during branch-and-bound and combines this *dependency analysis* approach with symbolic interval arithmetic and domain splitting techniques.

Moreover, both Tjeng et al. (2019) and Botoeva et al. (2020) report state-of-the-art performance on various network architectures and datasets but their tools consume very substantial amounts of CPU time. Depending on the classifier to be verified, we observed that finding a solution can easily take up to several hours of computation time for a single instance. Network verification can therefore turn into an extremely time-consuming

endeavour, even for a relatively small dataset, such as MNIST. At the same time, a verifier fails to maintain the premise of completeness, meaning that it can certify every input example it is presented with if many instances are subject to timeouts, which we also found to be the case for the verification methods considered in this study.

2.3 Automated algorithm configuration

Commercial tools for combinatorial problem solving usually come with many (hyper-)parameters, whose settings may have strong effects on the running time required for solving given problem instances. Deviating from the default and manually setting these performance parameters is a complex task that requires extensive domain knowledge and experimentation, and can be automated using algorithm configuration techniques.

In general, the algorithm configuration problem can be described as follows: Given an algorithm A (also referred to as the *target algorithm*) with parameter configuration space Θ, a set of problem instances Π and a cost metric $c : \Theta \times \Pi \rightarrow \mathbb{R}$, find a configuration $\theta^* \in \Theta$ that minimises cost c across the instances in Π:

$$\theta^* \in \arg\min_{\theta \in \Theta} \sum_{\pi \in \Pi} c(\theta, \pi) \tag{4}$$

The general workflow of the algorithm configuration procedure starts with picking a configuration $\theta \in \Theta$ and an instance $\pi \in \Pi$. Next, the configurator initialises a run of algorithm A with configuration θ on instance π with a maximal CPU time cutoff k and measures the resulting cost $c(\theta, \pi)$. The configurator uses this information about the target algorithm's performance to find a configuration that performs well on the training instances. Once its configuration budget (e.g., time budget) is exhausted, it returns its current incumbent θ^*, i.e., the best configuration found so far. Finally, when running the target algorithm with configuration θ^*, it should result in lower cost (such as average running time) across the benchmark set.

Automated algorithm configuration has been shown to work effectively in the context of SAT solving (Hutter et al., 2007, 2017), scheduling (Chiarandini et al., 2008), mixed-integer programming (Hutter et al., 2010; Lopez-Ibanez & Stützle, 2014), evolutionary algorithms (Bezerra et al., 2015), answer set solving (Gebser et al., 2011), AI planning (Vallati et al., 2013) and machine learning (Thornton et al., 2013; Feurer et al., 2015).

In this study, we use SMAC (Hutter et al., 2011), a widely known, freely available, state-of-the-art configurator based on sequential model-based optimisation (also known as Bayesian optimisation). The main idea of SMAC is to construct and iteratively update a statistical model of target algorithm performance (specifically: a random forest regressor; Breiman, 2001) to guide the search for good configurations. The random forest regressor allows SMAC to handle categorical parameters and therefore makes it suitable for MIP solvers, which have many configurable categorical parameters; SMAC has been shown to improve the performance of the commercial CPLEX solver over previous configuration approaches on several widely studied benchmarks (Hutter et al., 2011).

2.4 Portfolio construction

For the configuration procedure to work effectively, the problem instances of interest have to be sufficiently similar, such that a configuration that performs well on a subset of them also performs well on others. In other words, the instance set should be homogeneous. If a

given instance set does not satisfy this homogeneity assumption, automated configuration likely results in performance improvements on some instances, while performance on others might suffer, making it difficult to achieve overall performance improvements.

This problem can be addressed through *automatic portfolio construction* (Xu et al., 2010; Kadioglu et al., 2011; Malitsky et al., 2012; Lindauer et al., 2015). The general concept behind automatic portfolio construction techniques is to create a set of algorithm configurations that are chosen such that they complement each other's strengths and weaknesses. This portfolio should then be able to exploit per-instance variation much more effectively than a single algorithm configuration, which is designed to achieve high overall performance but may perform badly on certain types or subsets of instances.

More specifically, Hydra (Xu et al., 2010) automatically constructs portfolios containing multiple instances of the target algorithm with different configurations. The key idea behind Hydra is that a new candidate configuration is scored with its actual performance only in cases where it works better than any of the configurations in the existing portfolio, but with the portfolio's performance in cases where it performs worse. Thereby, a configuration is only rewarded to the extent that it improves overall portfolio performance and is not penalised for performing poorly on instances for which it should not be run anyway.

The portfolio construction procedure works as follows. Hydra starts with an initially empty portfolio $P := \{\}$ and executes several runs of target algorithm A. The configurator executes target algorithm A with different parameter configurations, searching for the algorithm configuration θ_i that yields the largest improvement in performance over P across the benchmark instances. Hydra evaluates the incumbent configurations returned from the configurator $\{\theta_1, \theta_2, ..., \theta_n\}$ and adds the k best to the portfolio: $P := P \cup \{\theta_1, \theta_2, ..., \theta_k\}$. Hydra then follows the same process in an iterative fashion, where the configurator finds new configurations to add to the portfolio at each iteration. The procedure terminates after a predefined set of iterations or after performance stagnates.

Once a portfolio has been constructed, there are essentially two ways to leverage the performance complementarity of the configurations contained in the portfolio. The first option is to extract instance-specific features and use those to train a statistical model that predicts the performance of each configuration in the portfolio individually. These predictions can then be used to select the configuration with the best-predicted performance (see, e.g., Xu et al, 2011). Alternatively, all configurations can be run in parallel on a given problem instance, which implicitly ensures that we always benefit from the best-performing configuration in the portfolio, at the cost of increased use of parallel resources. An empirical comparison between both approaches has been presented by Kashgarani and Kotthoff (2021).

3 Network verification with parallel MIP solver portfolios

In order to reduce complexity, Tjeng et al. (2019) mainly focused on reducing the number of variables in the verification problem. On the other hand, Botoeva et al. (2020) rely on pruning the search space during the branch-and-bound procedure. However, the embedded MIP solver and its numerous parameters were left untouched in both cases. More specifically, both methods employed a commercial MIP solver with default settings. This decision, along with their problem formulation, forms the starting point for our work.

More concretely, we seek to improve the performance of MIP-based neural network verification through configuring the MIP solver embedded in these systems, and constructing

a portfolio of solver configurations optimised for the benchmark set at hand; Fig. 1 provides an overview of the framework we propose. In brief, for a given network-example pair, we employ the verifier with several, differently configured instances of the embedded MIP solver. This portfolio of solvers is run in parallel and finishes once one solver has returned a solution or a global time limit has been reached.

In the following sections, we describe details of the configuration procedure as well as the MIP solver we configured.

3.1 Configuration procedure

In this work, we configure the commercial MIP solver Gurobi; see Sect. 3.2 for further details. Though it should be noted that, in principle, our approach works for any MIP solver.

The configuration procedure employs running Hydra over a predefined set of iterations to construct a portfolio of solver configurations with complementary strengths. The number of iterations is a hyper-parameter of the Hydra algorithm and has to be specified by the user. Since we cannot know the optimal portfolio size for a given benchmark in advance, we run Hydra over a reasonably larger number of iterations and, once the procedure has finished, discard configurations that did not improve portfolio performance on the validation set, i.e. that led to stagnation or reduction in total CPU time compared to the previous iteration. Note that the portfolio can contain the default configuration of the MIP solver.

Interestingly enough, we consistently observed strong heterogeneity among the instances in our benchmarks sets, making the use of a single configuration, i.e., a portfolio of size 1, ineffective. This is illustrated in Figure 2: Employing two different configurations individually on the same benchmark set shows that none of them outperforms the other, i.e., consistently achieves better performance across the entire set of instances. Combining both configurations into a portfolio, however, makes use of the complementary strengths of the configurations, and thereby achieves the highest overall performance, which motivates our choice of the portfolio approach.

Leveraging standard multi-core CPU architectures, we run the configurations in the portfolio in parallel until one of them returned a solution or until an overall limit on CPU time was exceeded. We note that, in principle, automated algorithm selection (see, e.g.,

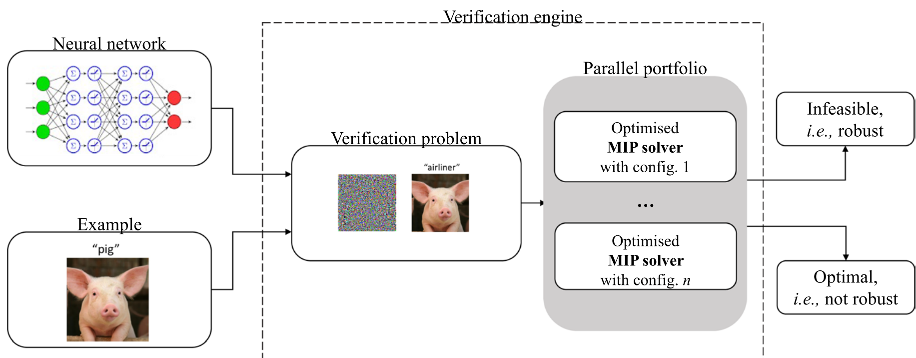

Fig. 1 Schematic diagram of the proposed framework

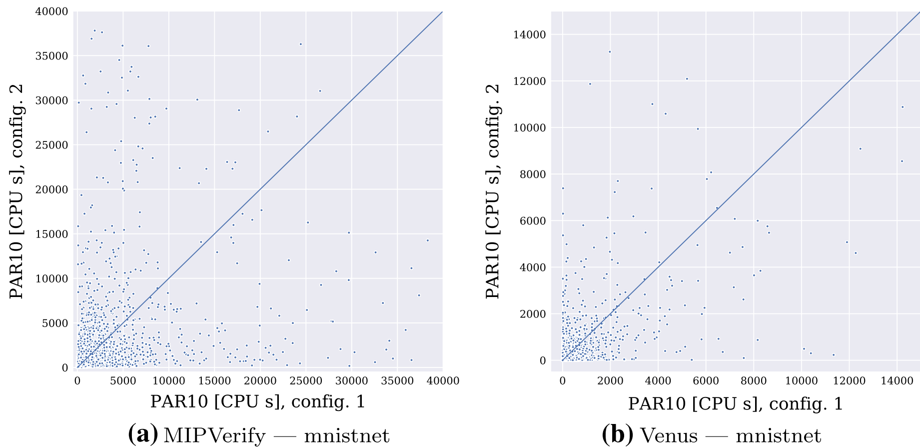

Fig. 2 Performance comparison of the configurations in the portfolios constructed for (a) MIPVerify and (b) Venus on the mnistnet benchmark. The plots show, that each configuration outperforms the other on some instances, while none of the configurations is dominating in performance across the entire benchmark set. This illustrates the complementary strengths of the configurations, which are exploited through portfolio construction. Note that there are also several instances on which one of the configurations reaches the time limit, but which are solved by the other. These are not shown in the figure due to the scaling of the axes. The diagonal line indicates equal performance of the two configurations

Kotthoff, 2016) could be used to determine from this portfolio the configuration likely to solve any given instance most efficiently, though this requires substantial amounts of training data and creates uncertainty from sub-optimal choices made by the machine learning technique at the heart of such selection approaches.

3.2 MIP solver

Following Tjeng et al. (2019) and Botoeva et al. (2020), we used the Gurobi MIP solver with a free academic license. Using the online documentation on Gurobi's parameters, we selected 62 performance-relevant parameters for configuration. These parameters can be categorical, e.g., the simplex variable pricing strategy parameter can take the values {Automatic (-1), Partial Pricing (0), Steepest Edge (1), Devex (2), and Quick-Start Steepest Edge (3)}, or continuous, e.g., the parameter controlling the magnitude of the simplex perturbation can take any value in the range $\{0, \infty\}$.

To control and limit the computational resources given to the solver, we fixed the number of CPU cores, i.e., the parameter *Threads*, to the value of 1. Thereby, we also ensure that the solver is optimised in such a way that it uses minimal computational resources, which, in turn, allows for more efficient parallelisation. In contrast, the default value of this parameter is an automatic setting, which means that the solver will generally use all available cores in a machine. There are further parameters that have an automatic setting as one of their values. In those cases, we allowed for the "automatic" value to be selected, but also other values.

While configuring the MIP solver embedded in MIPVerify is a rather straightforward task, additional considerations arise when configuring the solver embedded in Venus.

Essentially, Venus can run two modes, which lead to changes in the configuration space of the MIP solver: (*i*) Venus with *ideal cuts* and *dependency cuts* activated (default mode), in which case several cutting parameters are deactivated in Gurobi and therefore should be left untouched during the configuration procedure; (*ii*) Venus with its cutting mechanism deactivated, which allows for Gurobi's full parameter space to be optimised upon. Along with other, previously mentioned challenges, these considerations illustrate the complexity of adapting automated algorithm configuration techniques to the domain of neural network verification.

In order to maximally exploit the potential of automated hyperparameter optimisation, we decided to provide the configurator with full access to the configuration space and, thus, employ Venus with *ideal cuts* and *dependency cuts* deactivated and Gurobi's cutting parameters activated during portfolio construction.

4 Experimental setup

We test our method on several benchmarks, which will be introduced in the following, along with the objective of our configuration approach and the computational environment in which experiments were carried out.

4.1 Configuration objective

The objective of our configuration experiments is to minimise mean CPU time over all instances from the benchmark set. This choice deviates from the commonly used performance metric in the neural network verification literature, where evaluation is typically performed by operating on a fixed number of CPU cores while measuring wall-clock time. However, we do not consider wall-clock time a sensible performance measure when the evaluated methods use different numbers of cores. Instead, we decide to capture performance by means of CPU time, as it compensates for the possible difference in utilised cores. In other words, by choosing CPU time over wall-clock time, we ensure a more rigorous performance evaluation of our method as well as the baseline approaches, as one could easily gain performance in terms of wall-clock time through parallelisation, while heavily compromising in CPU time. Furthermore, we consider CPU time to be the more sensible performance measure, due to the cost associated with computational efforts. In fact, the rates for cloud services increase with the number of cores in a machine.

Generally, if the cost metric is running time, configurators typically optimise penalised average running time (PAR), notably PARk, as the metric of interest, which penalises unsuccessful runs by counting runs exceeding the cutoff time t_c as $t_c \times k$. In line with common practice in the algorithm configuration literature, we use $k = 10$ and refer to the cost metric as PAR10.

4.2 Details of the configuration procedure

The parameters for the configuration procedure were set as follows. Hydra ran over a predefined set of four iterations, during which it performed two independent runs of SMAC with a time budget of 24 hours each. Thus, running Hydra took $4 \times 2 \times 24 = 192$ hours for training, in addition to a variable amount of time spent on validation. In theory, the number of iterations could be set to a larger value; however, we refrained from

this to keep our experiments within reasonable time frames. Lastly, we set $k = 1$, which means that after every run, Hydra added one configuration to the portfolio, i.e., the configuration that yielded the largest gain in overall training performance. The final output, therefore, is a portfolio containing a minimum number of 1 and a maximum number of 4 solver configurations.

4.3 Data

Our benchmark sets were comprised of randomly chosen verification problem instances created by MIPVerify and Venus, respectively, using the network weights of two MNIST classifiers as well as the property-network pairs from the ACAS Xu repository (Julian et al., 2016; Katz et al., 2017). ACAS Xu contains an array of neural networks trained for horizontal manoeuvre advisory in unmanned aircraft. The MNIST classifiers were taken from the works of Tjeng et al. (2019) and Botoeva et al. (2020), respectively, and used to cross-test each verifier on both networks. The ACAS Xu benchmark was chosen to find out whether a high diversity in networks (the ACAS Xu repository contains 45 different neural networks) poses any challenges to the configuration procedure.

MNIST Firstly, we created problem instances using the network weights of the robust classifier $\text{SDP}_\text{d}\text{MLP}_\text{A}$ from Raghunathan et al. (2018). Among the networks considered in the work of Tjeng et al. (2019), we regard this one as the most difficult to verify, since it shows the largest average solving times and optimality gaps for many examples, even compared to classifiers trained on the typically more challenging CIFAR-10 benchmark. Secondly, we used the weights of the network mnistnet from the Venus repository (Botoeva et al., 2020), which is the only MNIST classifier considered in their study. In both cases, we created 184 instances, which were split 50-50 into disjoint training and validation sets. The training and validation sets were used during the configuration procedure, whereas the remaining 9 816 instances form the test set and were used to evaluate the final portfolio.

ACAS Xu For this benchmark, we only considered verification problem instances created by Venus, as MIPVerify at default reached the time limit of 38 400 CPU seconds for more than 80% of the instances. This makes automated configuration infeasible, as these instances do not only cause the default solver to time out but also any solver configuration tried by SMAC. Thereby, the configurator can hardly identify promising regions of the hyperparameter space and, consequently, not exploit them. Using Venus, we created 20 instances for different property-network pairs and, again, split them into disjoint training and validation sets. The remaining 152 instances are used for testing the final portfolio. Note that ACAS-Xu shows the highest average solving time among all benchmarks considered in the work of Botoeva et al. (2020).

4.4 Execution environment and software used

Our experiments were carried out on Intel Xeon E5-2683 CPUs with 32 cores, 40 MB cache size and 94 GB RAM, running CentOS Linux 7. We used MIPVerify version 0.2.3, Venus version 1.01, SMAC version 2.10.03, Hydra version 1.1 and the Gurobi solver version 9.0.1.

5 Results

We report empirical results for our new approach and each baseline in the form of (*i*) the fraction of timeouts; and (*ii*) bounds on adversarial error (the fraction of the dataset for which a valid adversarial example can be found), complement to adversarial accuracy (the fraction of the dataset known to be robust); (*iii*) CPU time (i.e., PAR10 scores) on solvable instances, i.e., instances that were solved by our portfolio or any of the baselines within the given cutoff time. Aggregated performance numbers are presented in Table 1 for MIPVerify and Table 2 for Venus, whereas Figs. 3 and 4 visualise penalised running time of our portfolio approach against the baselines on an instance level. Generally, we determined statistical significance using a binomial test with $\alpha = 0.05$ for timeouts and error bounds, and a permutation test with the number of permutations set at 10 000 and significance threshold of 0.05 for PAR10 scores.

5.1 MIPVerify

The results from our configuration experiments on the SDP_dMLP_A classifier are compared against multiple baselines. Firstly, we evaluated our portfolio approach against Gurobi, as used by Tjeng et al. (2019), using all 32 cores per CPU available on our compute cluster, with the cutoff time set to $1\,200 \times 32 = 38\,400$ CPU seconds (i.e., 1 200 seconds wall-clock time on a CPU without any additional load). In addition, since our parallel portfolio used 1 core for each of its 4 component configurations, we gathered additional baseline results from running the default configuration of Gurobi on the same number of cores and with the same cutoff as our portfolio, i.e., $9\,600 \times 4 = 38\,400$ CPU seconds. Lastly, to maximise the number of instances processed in parallel, we considered Gurobi in its default configuration limited to a single CPU core, with cutoff time of 38 400 seconds. In short, we

Table 1 Timeouts, adversarial error and PAR10 scores for different solver configurations of the MIP solver embedded in the MIPVerify engine on the MNIST dataset. Note that all approaches were given the same budget in terms of CPU time (the number of cores times the cutoff time)

Configuration	Cores	Cutoff [Seconds]	Timeouts	Adversarial error		PAR10 [CPU s]
				Lower Bound	Upper Bound	
SDP_dMLP_A classifier (Raghunathan et al., 2018)						
Default [SOTA]	32	1 200	21.29%	14.37%	30.67%	39 772
Default	4	9 600	17.74%	14.40%	27.49%	22 065
Default	1	38 400	17.66%	14.36%	27.58%	20 117
Portfolio [Ours]	4	9 600	**14.96**%	14.43%	**23.86**%	**8 478**
mnistnet classifier (Botoeva et al., 2020)						
Default	1	38 400	1.57%	69.96%	70.16%	2 969
Portfolio [Ours]	2	19 200	**1.38**%	70.13%	70.14%	**1 844**

Using our portfolio, we achieved better performance than the state-of-the-art (SOTA) method of Tjeng et al. (2019) as well as the default configuration of Gurobi using different numbers of cores. Boldfaced values indicate statistically significant improvements according to a binomial test with $\alpha = 0.05$ for timeouts and error bounds, and a permutation test with the number of permutations set at 10 000 and significance threshold of 0.05 for PAR10 scores

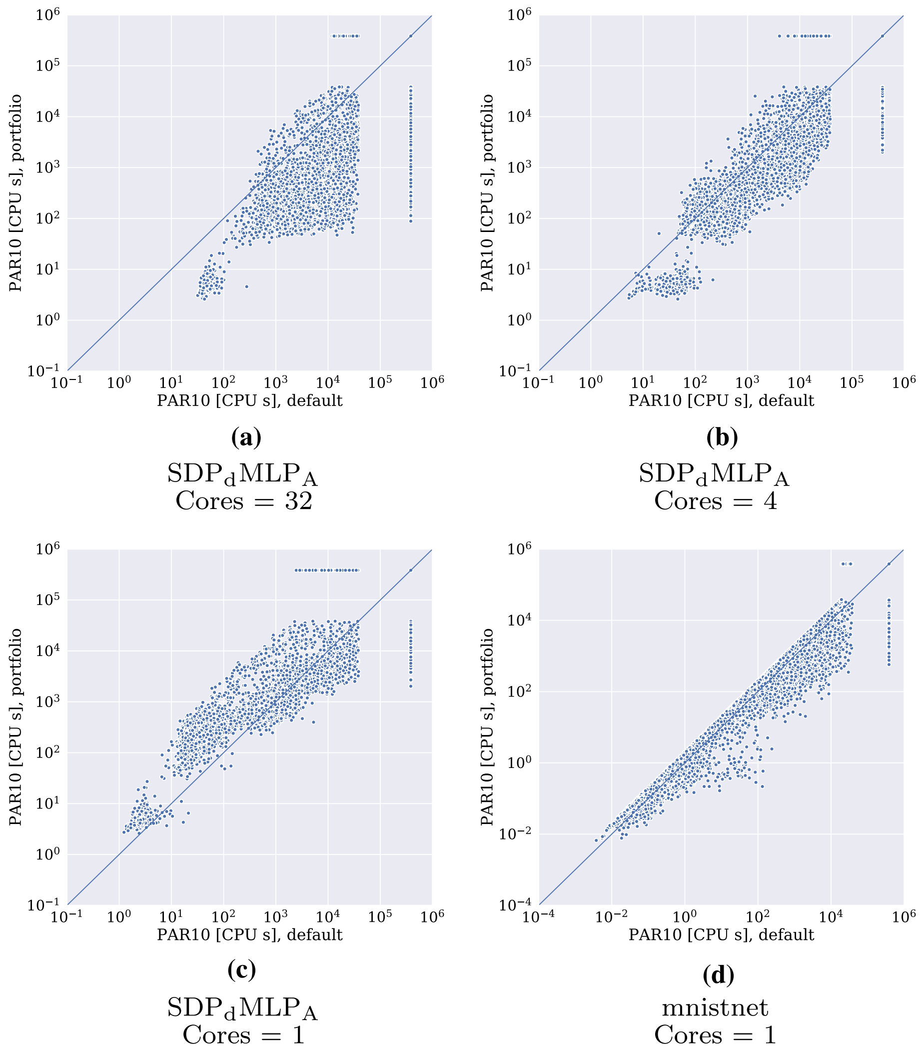

Fig. 3 Evaluation of our parallel portfolio approach for MIPVerify on the MNIST dataset (n=10 000) using weights from the SDP$_d$MLP$_a$ and mnistnet classifiers, respectively. Each dot represents a problem instance and the penalised running time for that instance achieved by the baseline approach (x-axis) vs our portfolio (y-axis). For SDP$_d$MLP$_a$, the baselines we considered are (a) the default solver running on all available, i.e., 32 cores, as in the work of Tjeng et al. (2019), (b) the default solver running on 4 cores and (c) the default solver running on 1 core. Our parallel portfolio, using 4 cores, achieved substantially fewer timeouts than any of the baselines and lower CPU times (in terms of PAR10 scores). Points grouped at the top and right border represent instances for which the solver reached the time limit, and are measured according to their penalised running time values

compared our approach against baselines with a variable number of cores and a constant budget in terms of CPU time. From these approaches, we considered only the best-performing one as the baseline for our configuration experiments on the mnistnet classifier.

As seen in Table 1, our portfolio was able to certify a statistically significantly larger fraction of instances, while reducing CPU time by an average factor of 4.7 on the

Table 2 Timeouts, adversarial error and PAR10 scores for different configurations of the MIP solver embedded in the Venus engine on the MNIST and ACAS Xu datasets

Configuration	Cores	Cutoff [Seconds]	Timeouts	Adversarial error Lower Bound	Adversarial error Upper Bound	PAR10 [CPU s]
mnistnet classifier (Botoeva et al., 2020)						
Default* [SOTA]	2	7 200	1.63%	70.33%	71.96%	1 975
Portfolio [Ours]	2	7 200	**0.58%**	70.61%	71.19%	**272**
SDP_dMLP_A classifier (Raghunathan et al., 2018)						
Default	1	14 400	9.76%	14.36%	24.12%	6 534
Portfolio [Ours]	2	7 200	**6.10%**	14.31%	**20.41%**	**636**
ACAS Xu (Julian et al., 2016; Katz et al., 2017)						
Default* [SOTA]	2	7 200	1.75%	20.34%	22.09%	1 314
Portfolio [Ours]	2	7 200	1.17%	20.34%	21.21%	**443**

Note that all approaches were given the same budget in terms of CPU time (the number of cores times the cutoff time). Using our portfolio, we achieved better performance than the state-of-the-art (SOTA) method of Botoeva et al. (2020). Boldfaced values indicate statistically significant improvements according to a binomial test with $\alpha = 0.05$ for timeouts and error bounds, and a permutation test with the number of permutations set at 10 000 and significance threshold of 0.05 for PAR10 scores. The asterisk marks Venus runs using the hyperparameter settings suggested by Botoeva et al. (2020), yet with Gurobi at default

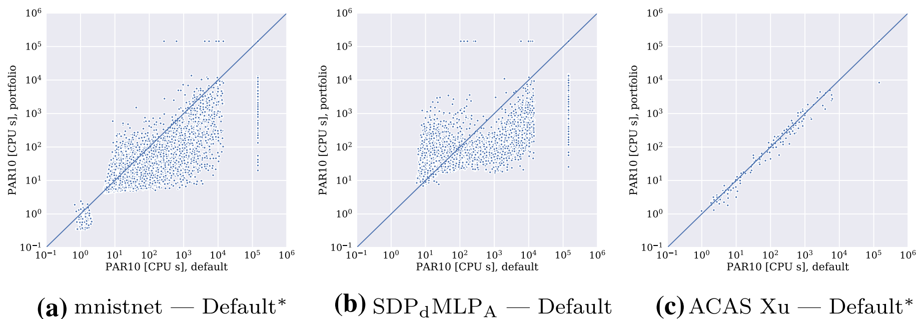

Fig. 4 Evaluation of our parallel portfolio approach for Venus on the MNIST dataset (n=10 000) using weights from the SDP_dMLP_a and mnistnet classifiers, respectively, and on the 172 property-network pairs from the ACAS Xu benchmark. Each dot represents a problem instance and the penalised running time for that instance achieved by the verifier with the embedded MIP solver at default (x-axis) vs our portfolio (y-axis). Overall, our parallel portfolio achieved fewer timeouts than the baseline and lower CPU times (in terms of PAR10 scores)

solvable instances (8 478 vs 39 772 CPU seconds). Furthermore, the portfolio strongly outperformed this baseline in terms of timeouts (14.96% vs 21.29%). More concretely, 694 instances solved by the portfolio timed out in the default setup with 32 cores; see Fig 3a for more details. 1 435 instances were neither solved by the default nor the portfolio within the given time limit. 61 instances on which the portfolio timed out were solved by the default solver.

The default configuration of Gurobi running on 4 cores was also clearly outperformed by our portfolio in terms of CPU time (8 478 vs 22 065 CPU seconds). Furthermore, the portfolio was able to reduce the number of timeouts (14.96% vs 17.74%), while improving on the upper bound (23.86% vs 27.49%). In other words, the portfolio certified more instances using fewer computational resources, although it was provided with the same number of cores and overall time budget. Figure 3b shows per-instance results for this set of experiments. Here, the default solver timed out on 378 instances, which were solved by the portfolio. On 109 instances, only the portfolio timed out. On 1 374 instances, both setups resulted in timeouts.

Lastly, we compared the portfolio against the default configuration of Gurobi running on a single-core. Here, our portfolio showed improved performance in terms of PAR10 (8 478 vs 20 117 CPU seconds) as well as the fraction of timeouts (14.96% vs 17.66%) and the upper bound (23.86% vs 27.58%). More specifically, the single-core default timed out on 378 instances that could be solved by the portfolio. On 108 instances, only the portfolio timed out. On 1 388 instances, both setups resulted in timeouts; see Fig 3c for more details.

On the mnistnet classifier, our portfolio also outperformed the single-core baseline in terms of PAR10 (1 844 vs 2 969 CPU seconds) as well as the fraction of timeouts (1.38% vs 1.57%), although to a smaller extent. To be precise, the default baseline timed out on 44 instances that the portfolio was able to solve (Fig 3d). On 25 instances, only the portfolio reached the time limit. 113 instances were neither solved by the default nor the portfolio. The default baseline timed out on 44 instances that the portfolio was able to solve. On 25 instances, only the portfolio reached the time limit. 113 instances were neither solved by the default nor the portfolio. These results could be explained by the mnistnet network being comparatively smaller and, thus, easier to verify than the $\mathrm{SDP_dMLP_A}$ classifier, as the latter results in a much larger number of timeouts when verified with equal settings.

5.2 Venus

The results from our configuration experiments are compared against two baseline approaches. Firstly, we evaluated our portfolio against Venus as employed by Botoeva et al. (2020), i.e., using the same hyperparameter settings for the verifier. We refer to this setup as *default**, as the MIP solver is left in its default configuration, while the verification engine is deployed with optimised hyperparameter settings. We note that the number of cores is equivalent to the number of parallel workers, which is set as a hyperparameter of the verifier. More precisely, we were running Venus using 2 workers, i.e., 2 cores per CPU available on our compute cluster, with the cutoff time set to $7\,200 \times 2 = 14\,400$ CPU seconds. In this setup, Venus employs 2 instances of the MIP solver in parallel, while we ensured that each solver is using exactly 1 CPU core. This way, we are giving the same amount of resources to the verifier and the portfolio. It should be noted that for the ACAS Xu benchmark, we also ran Venus with the hyperparameter settings reported by Botoeva et al. (2020), however with different numbers of workers. That is, we ran the verifier using 4 workers, 2 workers, and 1 worker, i.e., CPU core(s), to assess the effects of parallelism, and found CPU time to be constant with regards to the number of workers running in parallel. We, therefore, consider each of these baselines to be equally competitive and only report results for Venus running with 2 active workers, i.e., on 2 CPU cores and, thus, similar to the number of cores utilised by the portfolio.

As there is no optimal setting of Venus hyperparameters provided for the $\mathrm{SDP_dMLP_A}$ classifier, we used Venus with default settings as the baseline for our configuration

experiments on this benchmark. In this setup, Venus is running with 1 active worker, which uses the same overall time budget of 14 400 CPU seconds.

As Table 2 shows, the portfolio strongly outperformed Venus with default settings. On the mnistnet benchmark, it was able to certify a statistically significantly larger fraction of instances, while reducing CPU time by an average factor of 7.26 on the solvable instances (272 vs 1 975 CPU seconds). Furthermore, the portfolio strongly reduced the number of timeouts (1.63% vs 0.58%) on this benchmark. More specifically, the verifier timed out for 115 instances that were solved by the portfolio. On the other hand, the portfolio reached the time limit on 10 instances, which could be solved by the default. On 48 instances, both approaches resulted in timeouts; see Fig. 4a for more details.

This baseline was also used to evaluate our portfolio approach on the ACAS Xu benchmark and, as previously mentioned, employed the verifier using the same hyperparameter settings as reported by Botoeva et al. (2020), although with the number of workers or CPU cores fixed at 2. Essentially, the portfolio was able to slightly improve the number of timeouts and statistically significantly reduce CPU time by an average factor of 2.97 on the solvable instances (443 vs 1 314 CPU seconds). In concrete terms, the portfolio could solve 1 instance on which the default solver reached the time limit; see Fig. 4c. For clarification, we achieved comparable performance gains over Venus running with 4 workers in parallel (443 vs 1 337 CPU seconds) as well as Venus running with 1 worker (443 vs 1 306 CPU seconds).

On the SDP_dMLP_A benchmark, the default baseline, i.e., Venus with default settings, was outperformed by the portfolio in terms of PAR10 (636 vs 6 534 CPU seconds) as well as the fraction of timeouts (6.10% vs 9.76%). In this setup, the default timed out on 379 instances solved by the portfolio (Fig. 4b). On 15 instances, only the portfolio reached the time limit. 597 instances were neither solved by the default nor the portfolio. Lastly, the portfolio strongly improved on the upper bound (20.41% vs 24.12%), which overall clearly demonstrates the strength of the portfolio approach.

6 Conclusions and future work

In this study, we have, for the first time, demonstrated the effectiveness of automated algorithm configuration and portfolio construction in the context of neural network verification. Applying these techniques to neural network verification is by no means a trivial extension, due to the high running times and heterogeneity of the problem instances to be solved. In order to address this heterogeneity, we constructed a parallel portfolio of optimised MIP solver configurations with complementary strengths. Our method advises on the ideal number of configurations in the portfolio and can be used in combination with any MIP-based neural network verification system. We empirically evaluated our method on two recent, state-of-the-art MIP-based verification systems, MIPVerify and Venus.

Our results show that the portfolio approach can significantly reduce the CPU time required by these systems on various verification benchmarks, while reducing the number of timeouts and, thus, certifying a larger fraction of instances.

In more concrete terms, we strongly improved the performance of MIPVerify via speed-ups in CPU time by an average factor of 4.7 on the MNIST classifier SDP_dMLP_A from Raghunathan et al. (2018) and 1.61 on the MNIST classifier mnistnet from Botoeva et al. (2020). At the same time, we were able to lower the number of timeouts for both benchmarks and tighten previously reported bounds on adversarial error. For the

Venus verifier, we achieved even larger improvements, i.e., 10.3- and 7.26-fold reductions in average CPU time on the SDP_dMLP_A and mnistnet networks, respectively. Beyond that, we strengthened the performance of Venus on the ACAS Xu benchmark, attaining a 2.97-fold speedup in average CPU time. Overall, our results highlight the potential of employing MIP-based neural network verification systems with optimised solver configurations and demonstrate how our method can consistently improve neural network verifiers that make use of MIP solvers. At the same time, we note that our method is inherently dependent on the default performance of the verifier at hand. In other words, we acknowledge that this work alone cannot scale existing methods to network sizes that are completely beyond the capabilities of these methods. However, our approach can significantly improve the running time of the verifier on the benchmarks it is able to certify, and thus moves the boundary of network/input combinations accessible to the verifier.

We see several fruitful directions for future work. Firstly, we plan to explore the use of per-instance algorithm configuration techniques to further reduce the computational cost of our approach. While our parallel portfolio approach is robust and makes good use of parallel computing resources, judicious use of per-instance algorithm selection techniques could potentially save some computational costs. We note that this will require the development of grounded descriptive attributes (so-called meta-features) for neural network verification, which we consider an interesting research project in its own right.

The neural network verification systems we considered in this work have additional hyperparameters. While our current approach focuses on the hyperparameters of the internal MIP solver, in future work, we will also configure the hyperparameters at the verification level. Due to the potential impact that this has on the MIP formulation and therefore on the running time of a given instance, this poses specific challenges for the algorithm configuration methods we use.

Finally, the portfolios we construct consist of multiple configurations of the same verification engine. In principle, we could also consider heterogeneous portfolios that contain configurations of different verification engines, which could lead to further improvements in the state of the art in neural network verification, and ultimately make it possible to verify networks far beyond the sizes that can be handled by the methods we have introduced here.

Author Contributions MK has conducted the research presented in this manuscript. HH and JvR have regularly provided feedback on the work, contributed towards the interpretation of results, and have critically revised the whole. All authors approve the current version to be published and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Funding This research was partially supported by TAILOR, a project funded by EU Horizon 2020 research and innovation program under GA No. 952215.

Data Availability The data (i.e., MIP instances) used in our experiments was generated using the verifiers by Tjeng et al. (2019) and Botoeva et al. (2020), and and is available under https://github.com/marti-mcfly/nn-verification.

Code availability All code that was used for this research is available under https://github.com/marti-mcfly/nn-verification.

Declarations

Competing interests All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

Employment All authors declare that there is no recent, present, or anticipated employment by any organization that may gain or lose financially through publication of this manuscript.

Research involving human participants Not applicable: this research did not involve human participants, nor did it involve animals.

Consent for publication Not applicable: this research does not involve personal data, and publishing of this manuscript will not result in the disruption of any individual's privacy.

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

Akintunde, M., Lomuscio, A., Maganti, L., & Pirovano, E. (2018) Reachability analysis for neural agent-environment systems. In *Proceedings of The Sixteenth International Conference on Principles of Knowledge Representation and Reasoning* (KR2018)

Bastani, O., Ioannou, Y., Lampropoulos, L., Vytiniotis, D., Nori, A., & Criminisi, A. (2016). Measuring neural net robustness with constraints. In *Proceedings of the 30th Conference on Neural Information Processing Systems (NeurIPS 2016)*, pp 2613–2621

Bezerra, L. C., López-Ibánez, M., & Stützle, T. (2015). Automatic component-wise design of multiobjective evolutionary algorithms. *IEEE Transactions on Evolutionary Computation, 20*(3), 403–417.

Botoeva, E., Kouvaros, P., Kronqvist, J., Lomuscio, A., & Misener, R. (2020). Efficient verification of ReLU-based neural networks via dependency analysis. In *Proceedings of The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI20)* (pp. 3291–3299)

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Bunel, R .R ., Turkaslan, I., Torr, P., Kohli, P., & Mudigonda, P. K. (2018). A unified view of piecewise linear neural network verification. In *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, pp. 4790–4799

Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In *Proceedings of the 38th IEEE Symposium on Security and Privacy (IEEE S &P 2017)*, pp. 39–57

Carlini, N., Katz, G., Barrett, C., & Dill, D. L. (2017) Provably Minimally-Distorted Adversarial Examples. arXiv preprint arXiv:1709.10207

Chen, P. Y., Sharma, Y., Zhang, H., Yi, J., & Hsieh, C. J. (2018). Ead: Elastic-net attacks to deep neural networks via adversarial examples. In *Proceedings of The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI18)*

Cheng, C. H., Nührenberg, G., & Ruess , H. (2017). Maximum resilience of artificial neural networks. In *Proceedings of The 15th International Symposium on Automated Technology for Verification and Analysis (ATVA2017)*, pp. 251–268.

Chiarandini, M., Fawcett, C., & Hoos, H. H. (2008). A Modular Multiphase Heuristic Solver for Post Enrolment Course Timetabling. In *Proceedings of the 7th International Conference on the Practice and Theory of Automated Timetabling (PATAT 2008).*

Cohen, J., Rosenfeld, E., & Kolter, Z. (2019). Certified adversarial robustness via randomized smoothing. In *Proceedings of the Thirty-Sixth International Conference on Machine Learning (ICML2019)*, pp 1310–1320.

Dutta, S., Jha, S., Sankaranarayanan, S., & Tiwari, A. (2018) Output range analysis for deep neural networks. In *Proceedings of The Tenth NASA Formal Methods Symposium (NFM 2018)*, pp. 121–138.

Dvijotham, K., Stanforth, R., Gowal, S., Mann, T. A., & Kohli, P. (2018). A Dual Approach to Scalable Verification of Deep Networks. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2018)*, pp. 550–559.

Ehlers, R. (2017). Formal verification of piece-wise linear feed-forward neural networks. In *Proceedings of the 15th International Symposium on Automated Technology for Verification and Analysis (ATVA 2017)*, pp. 269–286.

Feurer, M., Springenberg, J. T., & Hutter, F. (2015). Initializing Bayesian hyperparameter optimization via meta-learning. In *Proceedings of The Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI15)*

Fischetti, M., & Jo, J. (2018). Deep neural networks and mixed integer linear optimization. *Constraints, 23*(3), 296–309.

Gebser, M., Kaminski, R., Kaufmann, B., Schaub, T., Schneider, M. T., & Ziller, S. (2011). A portfolio solver for answer set programming: Preliminary report. In *Proceedings of The Tenth International Conference on Logic Programming and Nonmonotonic Reasoning (LPNMR2019)*, pp. 352–357.

Gehr, T., Mirman, M., Drachsler-Cohen, D., Tsankov, P., Chaudhuri, S., & Vechev, M. (2018). AI2: Safety and robustness certification of neural networks with abstract interpretation. In *Proceedings of the 39th IEEE Symposium on Security and Privacy (IEEE S &P 2018)*, pp. 3–18.

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

Hutter, F., Babic, D., Hoos, H. H., & Hu, A. J. (2007). Boosting verification by automatic tuning of decision procedures. In *Formal Methods in Computer Aided Design (FMCAD'07)*, pp. 27–34

Hutter, F., Hoos, H. H., Leyton-Brown, K., & Stützle, T. (2009). ParamILS: An automatic algorithm configuration framework. *Journal of Artificial Intelligence Research, 36,* 267–306.

Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2010). Automated Configuration of Mixed Integer Programming Solvers. In *Proceedings of the 7th International Conference on Integration of Artificial Intelligence (AI) and Operations Research (OR) Techniques in Constraint Programming (CPAIOR 2010)*, pp. 186–202

Hutter, F., Hoos, H. H., Leyton-Brown, K. (2011). Sequential model-based optimization for general algorithm configuration. In *Proceedings of the 5th International Conference on Learning and Intelligent Optimization (LION 5)*, pp. 507–523

Hutter, F., Lindauer, M., Balint, A., Bayless, S., Hoos, H., & Leyton-Brown, K. (2017). The configurable SAT solver challenge (CSSC). *Artificial Intelligence, 243,* 1–25.

Julian, K. D., Lopez, J., Brush, J. S., Owen, M. P., & Kochenderfer, M. J. (2016). Policy compression for aircraft collision avoidance systems. In *Proceedings of the Thirty-Fifth Digital Avionics Systems Conference (DASC2016)*, pp. 1–10

Kadioglu, S., Malitsky, Y., Sabharwal, A., Samulowitz, H., & Sellmann, M. (2011). Algorithm selection and scheduling. In *Proceedings of the Seventeenth International Conference on Principles and Practice of Constraint Programming (CP2011)*, pp. 454–469

Kashgarani, H., & Kotthoff, L. (2021). Is algorithm selection worth it? Comparing selecting single algorithms and parallel execution. In *AAAI Workshop on Meta-Learning and MetaDL Challenge*, pp. 58–64.

Katz, G., Barrett, C., Dill, D. L., Julian, K., & Kochenderfer, M. J. (2017). Reluplex: An efficient SMT solver for verifying deep neural networks. In *Proceedings of the 29th International Conference on Computer Aided Verification(CAV 2017)*, pp. 97–117

König, M., Hoos, H. H., van Rijn, J. N. (2021). Speeding up neural network verification via automated algorithm configuration. In *ICLR Workshop on Security and Safety in Machine Learning Systems*.

Kotthoff, L. (2016). Algorithm selection for combinatorial search problems: A survey. In *Data Mining and Constraint Programming*. Springer, pp. 149–190.

Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533

Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., & Jana S (2019) Certified robustness to adversarial examples with differential privacy. In *Proceedings of The Fortieth IEEE Symposium on Security and Privacy (SP2019)*, IEEE, pp 656–672.

Lindauer, M., Hoos, H. H., Hutter, F., & Schaub, T. (2015). AutoFolio: An automatically configured algorithm selector. *Journal of Artificial Intelligence Research, 53,* 745–778.

Lomuscio, A., & Maganti, L. (2017). An approach to reachability analysis for feed-forward ReLU neural networks. arXiv preprint arXiv:1706.07351

Lopez-Ibanez, M., & Stützle, T. (2014). Automatically improving the anytime behaviour of optimisation algorithms. *European Journal of Operational Research, 235*(3), 569–582.

Malitsky, Y., Sabharwal, A., Samulowitz, H., & Sellmann, M. (2012). Parallel SAT Solver Selection and Scheduling. In *Proceedings of the Eighteenth International Conference on Principles and Practice of Constraint Programming (CP2012)*, pp. 512–526

Mohapatra, J., Ko, C. Y., Weng, L., Chen, P. Y., Liu, S., & Daniel, L. (2021). Hidden cost of randomized smoothing. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics (AISTATS2021)*, pp 4033–4041.

Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2016). Distillation as a defense to adversarial perturbations against deep neural networks. In *Proceedings of the 37th IEEE Symposium on Security and Privacy (IEEE S &P 2016)*, pp. 582–597.

Raghunathan, A., Steinhardt, J., & Liang, P. (2018). Certified defenses against adversarial examples. arXiv preprint arXiv:1801.09344

Scheibler, K., Winterer, L., Wimmer, R., & Becker, B. (2015). Towards verification of artificial neural networks. In *Proceedings of the 18th Workshop on Methoden und Beschreibungssprachen zur Modellierung und Verifikation von Schaltungen und Systemen (MBMV 2015)*, pp. 30–40.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., \& Fergus, R. (2014). Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199

Thornton, C., Hutter, F., Hoos, H. H., \& Leyton-Brown, K. (2013). *Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms*. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD2013)*, pp. 847–855

Tjeng, V., Xiao, .K, & Tedrake, R. (2019). Evaluating robustness of neural networks with mixed integer programming. In *Proceedings of the 7th International Conference on Learning Representations (ICLR 2019)*

Vallati, M., Fawcett, C., Gerevini, A. E., Hoos, H., \& Saetti, A. (2013). Automatic generation of efficient domain-specific planners from generic parametrized planners. In *Proceedings of the 6th Annual Symposium on Combinatorial Search (SOCS)*, pp. 184–192.

Wong, E., & Kolter, Z. (2018.) Provable defenses against adversarial examples via the convex outer adversarial polytope. In *Proceedings of The Thirty-Fifth International Conference on Machine Learning (ICML2018)*, pp 5286–5295.

Xiang, W., Tran, H. D., & Johnson, T. T. (2018). Output Reachable Set Estimation and Verification for Multilayer Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems, 29*(11), 5777–5783.

Xu L, Hoos H, Leyton-Brown K (2010) Hydra: Automatically Configuring Algorithms for Portfolio-Based Selection. In: Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI10)

Xu, L., Hutter, F., Hoos, H. H., Leyton-Brown, K. (2011). Hydra-MIP: Automated algorithm configuration and selection for mixed integer programming. In *RCRA Workshop on Experimental evaluation of Algorithms for Solving Problems with Combinatorial Explosion*, pp. 16–30

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