COMPLETE VERIFICATION VIA MULTI-NEURON RELAXATION GUIDED BRANCH-AND-BOUND

Anonymous authors
Paper under double-blind review

ABSTRACT

State-of-the-art neural network verifiers are fundamentally based on one of two paradigms: encoding the whole problem via tight multi-neuron convex relaxations or applying a Branch-and-Bound (BaB) procedure which leverages imprecise but fast bounds on a large number of easier subproblems. The former can better capture complex multi-neuron dependencies but sacrifices completeness due to inherent precision limits of convex relaxations. The latter enables complete verification yet becomes increasingly ineffective on larger and more challenging networks. In this work, we present a novel complete verifier which combines the strengths of both paradigms: it leverages multi-neuron relaxations and an efficient, GPU-based dual optimizer to drastically reduce the number of subproblems generated during the BaB process. An extensive evaluation demonstrates that our verifier achieves new state-of-the-art results on both established benchmarks as well as networks with significantly higher accuracy than previously considered. The latter result (up to 26% certification gains) indicates meaningful progress towards creating verifiers that can handle practically relevant networks.

1 INTRODUCTION

Recent years have witnessed substantial interest in new methods for certifying properties of neural networks, ranging from stochastic methods (Cohen et al., 2019) which construct a robust model from an underlying classifier to deterministic methods (Gehr et al., 2018; Katz et al., 2017; Xu et al., 2020a) that analyze a given network as is (the focus of our work).

Key challenge: Scalable and Precise Non-Linearity Handling Deterministic verification methods can be categorized as complete or incomplete. Recent incomplete verification methods based on propagating and refining a single convex region (Müller et al., 2021; Dathathri et al., 2020; Tjandraatmadja et al., 2020) are limited in precision due to fundamental constraints imposed by convex relaxations. Traditional complete verification approaches based on SMT solvers (Ehlers, 2017) or a single mixed-integer linear programming encoding of a property (Tjeng et al., 2017; Katz et al., 2017) suffer from worst-case exponential complexity and are often unable to compute sound bounds in reasonable time-frames. To address this issue, a promising recent formulation in the form of a Branch-and-Bound approach (Bunel et al., 2020) has been popularized, yielding anytime-valid bounds, by recursively splitting the problem domain into easier subproblems and deriving bounds on each of these via cheap and less precise methods (Xu et al., 2020b; Wang et al., 2021; Palma et al., 2021; Henriksen & Lomuscio, 2021). This approach has proven effective on (smaller) networks where there are relatively few unstable activations and where splitting can yield substantial improvements. However, for larger networks or those not regularized to be amenable to certification, this strategy becomes increasingly ineffective as the larger number of unstable activations make individual splits less effective (exacerbated by the relatively loose underlying bounding methods).

This work: Branch-and-Bound guided by Multi-Neuron Constraints In this work, we propose a new certification method and verifier, called **M**ulti-**N**euron Constraint Guided **BaB** (MN-BAB), which aims to combine the best of both worlds: it builds on the tight multi-neuron constraints proposed by Müller et al. (2021) and leverages these constraints within a BaB framework to yield an efficient, GPU based dual solver. The key benefit here lies in the fact that exploiting the significantly increased precision of the underlying bounding method substantially reduces the number of domain

splits (carrying exponential cost) required to certify a property. This benefit is especially pronounced for larger and less regularized networks where additional splits of the problem domain yield diminishing returns. We release all code and scripts to reproduce our experiments at <ANONYMIZED>.

Main Contributions:

- We present a novel verification framework, MN-BAB, which leverages tight multi-neuron constraints and a GPU-based dual solver in a BaB approach.
- We develop a novel branching heuristic based on information obtained from analyzing our multi-neuron constraints.
- We propose a new class of branching heuristics, applicable to all BaB-based verifiers, that correct the expected bound improvement of a branching decision for the incurred computational cost.
- Our extensive empirical evaluation demonstrates that we improve on the state-of-the-art in terms of certified accuracy by as much as 26% while being 30% faster.

2 BACKGROUND

In this section, we review the necessary background underlying our method (discussed next).

2.1 NEURAL NETWORK VERIFICATION

The neural network verification problem can be defined as follows: given an input domain $\mathcal{D} \subseteq \mathcal{X}$, a network $f : \mathcal{X} \to \mathcal{Y}$, and a linear property $\mathcal{P} \subseteq \mathcal{Y}$ over the output neurons $y \in \mathcal{Y}$, prove $f(x) \in \mathcal{P}$, $\forall \boldsymbol{x} \in \mathcal{D}$. We instantiate this problem with the challenging ℓ_∞-norm bounded specifications and set $\mathcal{D}$ to the ℓ_∞ ball around an input point $\boldsymbol{x}_0$ of radius ϵ: $\mathcal{D}_\epsilon(\boldsymbol{x}_0) = \{\boldsymbol{x} \mid ||\boldsymbol{x} - \boldsymbol{x}_0||_\infty \leq \epsilon\}$.

For ease of presentation, we consider neural networks of L fully connected layers with ReLU activation functions (we note that MN-BAB can handle a wide range of layers including convolutional, residual, batch-normalization, average-pooling layers). We focus on ReLU networks as the BaB framework only yields complete verifiers for piecewise linear activation functions, but remark that our approach is applicable to a wide class of activations including ReLU, Sigmoid, Tanh, MaxPool, and others. We denote the number of neurons in the i^{th} layer as d_i and the corresponding weights and biases as $\boldsymbol{W}^{(i)} \in \mathbb{R}^{d_i \times d_{i-1}}$ and $\boldsymbol{b}^{(i)} \in \mathbb{R}^{d_i}$ for $i \in \{1, ..., L\}$. Further, the neural network is defined as $f(\boldsymbol{x}) := \hat{\boldsymbol{z}}^{(L)}(\boldsymbol{x})$ where $\hat{\boldsymbol{z}}^{(i)}(\boldsymbol{x}) := \boldsymbol{W}^{(i)}\boldsymbol{z}^{(i-1)}(\boldsymbol{x}) + \boldsymbol{b}^{(i)}$ and $\boldsymbol{z}^{(i)}(\boldsymbol{x}) := \max(0, \hat{\boldsymbol{z}}^{(i)}(\boldsymbol{x}))$. For readability, the notation omits the dependency of intermediate activations on $\boldsymbol{x}$.

Since we can encode any linear property over output neurons into an additional affine layer, we can simplify the general formulation $\boldsymbol{f}(\boldsymbol{x}) \in \mathcal{P}$ to $\boldsymbol{f}(\boldsymbol{x}) > \boldsymbol{0}$. The property can now be verified by proving that a lower bound is greater 0 via the following optimization problem:

$$\begin{aligned} \min_{\boldsymbol{x} \in \mathcal{D}_\epsilon(\boldsymbol{x}_0)} \quad & \boldsymbol{f}(\boldsymbol{x}) = \hat{\boldsymbol{z}}^{(L)} \\ s.t. \quad & \hat{\boldsymbol{z}}^{(i)} = \boldsymbol{W}^{(i)}\boldsymbol{z}^{(i-1)} + \boldsymbol{b}^{(i)} \\ & \boldsymbol{z}^{(i)} = \max(0, \hat{\boldsymbol{z}}^{(i)}) \end{aligned} \tag{1}$$

A method is called complete if it can prove every property that actually holds (no false negatives).

2.2 BRANCH-AND-BOUND FOR VERIFICATION

Bunel et al. (2020) successfully applied the Branch-and-Bound (BaB) approach (Land & Doig, 1960) to neural network verification. It consists of a bounding method that computes sound upper and lower bounds on the optimization objective of Eq. (1) and a branching method that recursively splits the problem into subproblems with added constraints, allowing for increasingly tighter bounds. If an upper bound < 0 is found, this represents an adversarial example and allows to terminate the procedure. If the obtained lower bound is > 0, the property is verified on the corresponding domain. If a lower bound > 0 is derived on all splits, the property is verified. An ideal splitting procedure minimizes the total time required for bounding, which is often well approximated by the number of

considered subproblems. A simple approach is to split the input domain, however, this is inefficient for high-dimensional input spaces. Splitting a ReLU activation node into its positive and negative phases has been shown to be far more efficient (Bunel et al., 2020) and ultimately yields a complete verifier (Wang et al., 2021). Hence, we focus solely on ReLU branching strategies.

2.3 Linear Constraints

The key challenge in neural network verification Eq. (1) are the non-linear activations. Every ReLU activation which can be either active ($\hat{z} >= 0$) or inactive ($\hat{z} < 0$) depending on the exact $x \in \mathcal{D}$ is called unstable and has to be approximated using a convex relaxation of its input-output-set. Stable ReLUs, in contrast, can simply be replaced by a linear function for the input set $\mathcal{D}$. We build on the convex relaxation introduced in DeepPoly (Singh et al., 2019b) and shown in Fig. 1. Its key property is the single linear upper and lower bound, which allows for efficient bound computation.

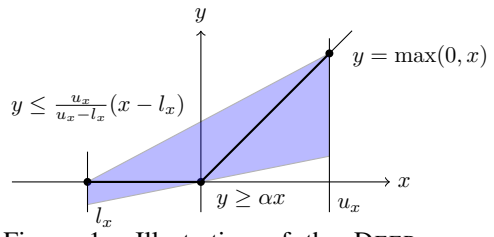

Figure 1: Illustration of the DeepPoly relaxation of a ReLU activation $y = \max(x, 0)$ given the neuron-wise bounds $x \in [l_x, u_x]$ and parametrized by $\alpha \in [0, 1]$.

2.4 Multi-Neuron Constraints

All convex relaxations that consider ReLU neurons individually are fundamentally limited in their precision by the so-called (single-neuron) convex relaxation barrier (Salman et al., 2019). It can be overcome by considering multiple neurons jointly (Singh et al., 2019a; Tjandraatmadja et al., 2020; Müller et al., 2021; Palma et al., 2021) capturing interactions between these neurons to obtain tighter bounds, illustrated in Fig. 2. We leverage the multi-neuron constraints from Müller et al. (2021), expressed as a conjunction of linear constraints over the joint input and output space of a ReLU layer.

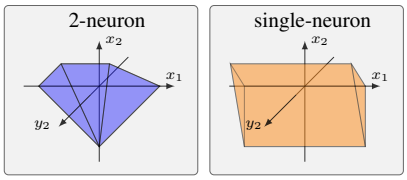

Figure 2: Comparison of multi-neuron and single-neuron constraints projected into y_2-x_1-x_2-space. Reproduced from Müller et al. (2021).

2.5 Constrained Optimization via Lagrange Multipliers

To incorporate constraints into an optimization problem we use the technique of Lagrange multipliers. Given a constrained minimization problem $\min_x f(x), s.t.\ g(x) \leq 0$, we can bound the objective function with:

$$\min_x f(x) \geq \min_x \max_{\lambda \geq 0} f(x) + \lambda g(x)$$

If the constraint is satisfied strictly, i.e., $g(x) < 0$, we set $\lambda = 0$ to maximize the objective, else, i.e., $g(x) \geq 0$, increasing λ would allow the objective to grow unboundedly, shifting the minimum over x until the constraint is satisfied. Hence, if $\lambda > 0$ after optimization, the constraint is active, i.e., it is actively enforced and currently satisfied with equality.

3 A Multi-Neuron Relaxation Based BaB Framework

In this section, we describe the two key components of MN-BaB: (i) an efficient bounding method leveraging multi-neuron constraints, and (ii) a branching method tailored to it.

3.1 Efficient Multi-Neuron Bounding

We build on the approach of Singh et al. (2019b), extended by Wang et al. (2021) of deriving a lower bound $\underline{f}$ as a function of the network inputs and a set of optimizable parameters. Crucially, we tighten these relatively loose bounds significantly by enforcing precise multi-neuron constraints via Lagrange multipliers. To enable this, we develop a method capable of integrating any linear constraint over arbitrary neurons anywhere in the network into the bounding objective. At a high level, we derive linear upper and lower bounds of the form $\boldsymbol{z}^{(i)} \geq \boldsymbol{A}\boldsymbol{z}^{(i-1)} + c$ for every layer's

output in terms of its inputs $\boldsymbol{z}^{(i-1)}$. Then, starting with a linear expression in the last layer's outputs $\boldsymbol{z}^{(L)}$ that we aim to bound, we use the linear bounds derived above, to replace $\boldsymbol{z}^{(L)}$ with symbolic bounds depending only on the previous layer's values $\boldsymbol{z}^{(L-1)}$. We proceed in this manner recursively, until we obtain an expression in terms of only the networks inputs. Below we describe this backsubstitution for ReLU and affine layers.

Affine Layer Assume any affine layer $\hat{\boldsymbol{z}}^{(i)} = \mathbf{W}^{(i)}\boldsymbol{z}^{(i-1)} + \mathbf{b}^{(i)}$ and a lower bound $\underline{f} = \hat{\boldsymbol{a}}^{(i)}\hat{\boldsymbol{z}}^{(i)} + \hat{c}^{(i)}$ with respect to its outputs. We then substitute the affine expression for $\hat{\boldsymbol{z}}^{(i)}$ to obtain:

$$\underline{f} = \underbrace{\hat{\boldsymbol{a}}^{(i)}\mathbf{W}^{(i)}}_{\boldsymbol{a}^{(i-1)}}\boldsymbol{z}^{(i-1)} + \underbrace{\hat{\boldsymbol{a}}^{(i)}\mathbf{b}^{(i)} + \hat{c}^{(i)}}_{c^{(i-1)}} = \boldsymbol{a}^{(i-1)}\boldsymbol{z}^{(i-1)} + c^{(i-1)} \tag{2}$$

ReLU Layer Let $\underline{f} = \boldsymbol{a}^{(i)}\boldsymbol{z}^{(i)} + c^{(i)}$ be a lower bound with respect to the output of a ReLU layer $\boldsymbol{z}^{(i)} = \max(0, \hat{\boldsymbol{z}}^{(i)})$ and $\mathbf{l}^{(i)}$ and $\mathbf{u}^{(i)}$ bounds on its input s.t. $\mathbf{l}^{(i)} \leq \hat{\boldsymbol{z}}^{(i)} \leq \mathbf{u}^{(i)}$, obtained by recursively applying the bounding procedure or using a cheaper but less precise propagation-based bounding method (Singh et al., 2018; Gowal et al., 2018). The backsubstitution through a ReLU layer now consists of three distinct steps: 1) enforcing multi-neuron constraints, 2) enforcing neuron-wise relaxations, and 3) enforcing split constraints, which we describe in detail below.

Enforcing multi-neuron constraints We compute multi-neuron constraints as described in Müller et al. (2021), although our approach is applicable to any linear constraints in the input-output space of ReLU activations, written as:

$$\begin{bmatrix}\boldsymbol{P}^{(i)} & \hat{\boldsymbol{P}}^{(i)} & -\mathbf{p}^{(i)}\end{bmatrix}\begin{bmatrix}\boldsymbol{z}^{(i)} \\ \hat{\boldsymbol{z}}^{(i)} \\ 1\end{bmatrix} \leq 0. \tag{3}$$

where $\hat{\boldsymbol{z}}^{(i)}$ are the pre- and $\boldsymbol{z}^{(i)}$ the post-activation values and $\boldsymbol{P}^{(i)}$, $\hat{\boldsymbol{P}}^{(i)}$, and $-\mathbf{p}^{(i)}$ the constraint parameters. We enforce these constraints using Lagrange multipliers (see §2.5), yielding sound lower bounds for all $\boldsymbol{\gamma} \in (\mathbb{R}^{\geq 0})^{e_i}$, where e_i denotes the number of multi-neuron constraints in layer i.

$$\begin{aligned}\boldsymbol{a}^{(i)}\boldsymbol{z}^{(i)} + c^{(i)} &\geq \max_{\boldsymbol{\gamma}^{(i)} \geq 0} \boldsymbol{a}^{(i)}\boldsymbol{z}^{(i)} + c^{(i)} + \boldsymbol{\gamma}^{(i)\top}(\boldsymbol{P}^{(i)}\boldsymbol{z}^{(i)} + \hat{\boldsymbol{P}}^{(i)}\hat{\boldsymbol{z}}^{(i)} + -\mathbf{p}_i) \\ &= \max_{\boldsymbol{\gamma}^{(i)} \geq 0} \underbrace{(\boldsymbol{a}^{(i)} + \boldsymbol{\gamma}^{(i)\top}\boldsymbol{P}^{(i)})}_{\boldsymbol{a}'^{(i)}}\boldsymbol{z}^{(i)} + \underbrace{\boldsymbol{\gamma}^{(i)\top}\hat{\boldsymbol{P}}^{(i)}\hat{\boldsymbol{z}}^{(i)} + \boldsymbol{\gamma}^{(i)\top}(-\mathbf{p}_i) + c^{(i)}}_{c'^{(i)}}\end{aligned}$$

Note that this approach can be easily extended to linear constraints over any activations in arbitrary layers if applied in the last affine layer at the very beginning of the backsubstitution process.

Enforcing single-neuron constraints We now apply the single-neuron DEEPPOLY relaxation with parametrized slopes α collected in $\boldsymbol{D}$ (see below):

$$\begin{aligned}\max_{\boldsymbol{\gamma}^{(i)} \geq 0} \boldsymbol{a}'^{(i)}\boldsymbol{z}^{(i)} + c'^{(i)} &\geq \max_{\substack{0 \leq \boldsymbol{\alpha}^{(i)} \leq 1 \\ \boldsymbol{\gamma}^{(i)} \geq 0}} \boldsymbol{a}'^{(i)}(\boldsymbol{D}^{(i)}\hat{\boldsymbol{z}}^{(i)} + \underline{b}^{(i)}) + c'^{(i)} \\ &= \max_{\substack{0 \leq \boldsymbol{\alpha}^{(i)} \leq 1 \\ \boldsymbol{\gamma}^{(i)} \geq 0}} \underbrace{\boldsymbol{a}'^{(i)}\boldsymbol{D}^{(i)}}_{\boldsymbol{a}''^{(i)}}\hat{\boldsymbol{z}}^{(i)} + \underbrace{\boldsymbol{a}'^{(i)}\underline{b}^{(i)} + c'^{(i)}}_{c''^{(i)}}\end{aligned}$$

The intercept vector $\underline{\boldsymbol{b}}$ and the diagonal slope matrix $\boldsymbol{D}$ are defined as:

$$D_{j,j} = \begin{cases} 1 & \text{if } l_j \geq 0 \text{ or node } j \text{ is positively split} \\ 0 & \text{if } u_j \leq 0 \text{ or node } j \text{ is negatively split} \\ \alpha_j & \text{if } l_j < 0 < u_j \text{ and } a_j \geq 0 \\ \frac{u_j}{u_j - u_j} & \text{if } l_j < 0 < u_j \text{ and } a_j < 0 \end{cases}$$

$$\underline{b}_j = \begin{cases} -\frac{u_j l_j}{u_j - l_j} & \text{if } l_j < 0 < u_j \text{ and } a_j < 0 \\ 0 & \text{otherwise} \end{cases}$$

Where we drop the layer index i for readability and α_j is the lower bound slope parameter illustrated in Fig. 1. Note how, depending on whether the sensitivity $a_j^{(i)}$ of $\underline{f}$ w.r.t. $z_j^{(i)}$ has positive or negative sign, we substitute $z_j^{(i)}$ for its lower or upper bound, respectively.

Enforcing split constraitns We encode split constraints of the form $\hat{z}_j^{(i)} \geq 0$ or $\hat{z}_j^{(i)} \leq 0$ using the diagonal split matrix $\boldsymbol{S}$ as follows, again dropping the layer index i:

$$S_{j,j} = \begin{cases} -1 & \text{positive split: } \hat{z}_j \geq 0 \\ 1 & \text{negative split: } \hat{z}_j < 0 \\ 0 & \text{no split} \end{cases}$$

$$\boldsymbol{S}^{(i)} \hat{\boldsymbol{z}}^{(i)} \leq 0$$

We again enforce these constraints using Lagrange multipliers:

$$\max_{\substack{0 \leq \boldsymbol{\alpha}^{(i)} \leq 1 \\ \boldsymbol{\gamma}^{(i)} \geq 0}} \boldsymbol{a}''^{(i)} \hat{\boldsymbol{z}}^{(i)} + c''^{(i)} \geq \max_{\substack{0 \leq \boldsymbol{\alpha}^{(i)} \leq 1 \\ \boldsymbol{\beta}^{(i)} \geq 0 \\ \boldsymbol{\gamma}^{(i)} \geq 0}} \underbrace{(\boldsymbol{a}''^{(i)} + \boldsymbol{\beta}^{(i)\top} \boldsymbol{S}^{(i)})}_{\boldsymbol{a}'''^{(i)}} \hat{\boldsymbol{z}}^{(i)} + \underbrace{\underline{c}''^{(i)}}_{c'''^{(i)}}$$

Putting everything together, the backsubstitution operation through a ReLU layer is:

$$\begin{aligned} \min_{x \in \mathcal{D}} \boldsymbol{a}^{(i)} \boldsymbol{z}^{(i)} + c^{(i)} \geq \min_{x \in \mathcal{D}} \max_{\substack{0 \leq \boldsymbol{\alpha}^{(i)} \leq 1 \\ \boldsymbol{\beta}^{(i)} \geq 0 \\ \boldsymbol{\gamma}^{(i)} \geq 0}} & \underbrace{((\boldsymbol{a}^{(i)} + \boldsymbol{\gamma}^{(i)\top} \boldsymbol{P}^{(i)}) \boldsymbol{D}^{(i)} + \boldsymbol{\beta}^{(i)\top} \boldsymbol{S}^{(i)} + \boldsymbol{\gamma}^{(i)\top} \hat{\boldsymbol{P}}^{(i)})}_{\hat{\boldsymbol{a}}^{(i)}} \hat{\boldsymbol{z}}^{(i)} \\ & + \underbrace{(\boldsymbol{a}^{(i)} + \boldsymbol{\gamma}^{(i)\top} \boldsymbol{P}^{(i)})' \underline{b}^{(i)} + \boldsymbol{\gamma}^{(i)\top} \text{-}\mathbf{p}^{(i)} + c^{(i)}}_{\hat{c}^{(i)}} \end{aligned} \tag{4}$$

Full backsubstitution through all layers leads to an optimizable lower bound on $\underline{f}$:

$$\min_{\boldsymbol{x} \in \mathcal{D}} f(x) \geq \min_{\boldsymbol{x} \in \mathcal{D}} \max_{\substack{0 \leq \boldsymbol{\alpha} \leq 1 \\ 0 \leq \boldsymbol{\beta} \\ 0 \leq \boldsymbol{\gamma}}} \boldsymbol{a}^{(0)} \boldsymbol{x} + c^{(0)} \geq \max_{\substack{0 \leq \boldsymbol{\alpha} \leq 1 \\ 0 \leq \boldsymbol{\beta} \\ 0 \leq \boldsymbol{\gamma}}} \min_{\boldsymbol{x} \in \mathcal{D}} \boldsymbol{a}^{(0)} \boldsymbol{x} + c^{(0)}$$

where the second inequality holds due to weak duality. We denote all $\alpha_j^{(i)}$ from every layer of the backsubstitution process with $\boldsymbol{\alpha}$ and define $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ analogously. The inner minimization over $\boldsymbol{x}$ has a closed form solution if $\mathcal{D}$ is an l_p-ball of radius ϵ around $\boldsymbol{x}_0$, given by Hölder's inequality:

$$\max_{\substack{0 \leq \boldsymbol{\alpha} \leq 1 \\ 0 \leq \boldsymbol{\beta} \\ 0 \leq \boldsymbol{\gamma}}} \min_{\boldsymbol{x} \in \mathcal{D}} \boldsymbol{a}^{(0)} \boldsymbol{x} + c^{(0)} \geq \max_{\substack{0 \leq \boldsymbol{\alpha} \leq 1 \\ 0 \leq \boldsymbol{\beta} \\ 0 \leq \boldsymbol{\gamma}}} \boldsymbol{a}^{(0)} \boldsymbol{x}_0 - ||\boldsymbol{a}^{(0)\top}||_q \, \epsilon + b^{(0)} \tag{5}$$

where q is defined s.t. $\frac{1}{p} + \frac{1}{q} = 1$. Since these bounds are sound for any $0 \leq \boldsymbol{\alpha} \leq 1$, and $\boldsymbol{\beta}, \boldsymbol{\gamma} \geq 0$, they can be tightened by using (projected) gradient ascent to optimize these parameters.

We compute all intermediate bounds using the same bounding procedure, leading to two full parameter sets for every neuron in the network. To reduce memory requirements, we share parameter sets between all neurons in the same layer, but keep separate sets for upper and lower bounds.

Upper Bounding the Minimum Objective Showing an upper bound on the minimum optimization objective precluding verification ($\underline{f} < 0$) allows us to terminate the BaB process early. Propagating any input $\boldsymbol{x} \in \mathcal{D}$ through the network yields a valid upper bound, hence, we propose to use the input that minimizes Eq. (5):

$$x_i = \begin{cases} (x_0)_i + \epsilon & \text{if } a_i^{(0)} < 0 \\ (x_0)_i - \epsilon & \text{if } a_i^{(0)} \geq 0 \end{cases}.$$

3.2 MULTI-NEURON CONSTRAINT GUIDED BRANCHING

The general BaB approach is based on recursively splitting an optimization problem into easier subproblems to derive increasingly tighter bounds. However, the benefit of different splits can vary widely, making an effective branching heuristic choosing splits that minimize the number of required subproblems a key component of any BaB framework (Bunel et al., 2020). Typically, a score is assigned based on the expected bound improvement and the most promising split is chosen. Consequently, the better the actual bound improvement is captured, the better the resulting decision. Both the commonly used BABSR Bunel et al. (2020) and the more recent FSB De Palma et al. (2021) approximate improvements under a DEEPPOLY style backsubstitution procedure. As they do not consider the impact of multi-neuron constraints at all, the scores they compute might become very noisy proxies for the bound improvements obtained with our method. To overcome this issue, we propose a novel branching heuristic, Active Constraint Score Branching (ACS), tailored to the multi-neuron constraints. Further, we introduce a branching heuristic framework which corrects the expected bound improvement with the potentially significantly varying expected computational cost.

Active Constraint Score Branching The value of a Lagrange parameter γ provides valuable information about the constraint it enforces. Concretely, $\gamma > 0$ indicates that a constraint is active, i.e., the optimal solution fulfills it with equality. Further, for a constraint $g(x) \leq 0$, a larger $\partial_x \gamma g(x)$ indicates a larger sensitivity of the final bound to violations of this constraint. We compute this sensitivity for our multi-neuron constraints with respect to both ReLU outputs and inputs as $\boldsymbol{\gamma}^\top \boldsymbol{P}$ and $\boldsymbol{\gamma}^\top \hat{\boldsymbol{P}}$, respectively, where $\boldsymbol{P}$ and $\hat{\boldsymbol{P}}$ are the multi-neuron constraint parameters. We then define our branching score for a neuron j in layer i as the sum over all sensitivities with respect to its input or output:

$$s_{i,j} = |\boldsymbol{\gamma}^{(i)\top} \boldsymbol{P}^{(i)}|_j + |\boldsymbol{\gamma}^{(i)\top} \hat{\boldsymbol{P}}^{(i)}|_j, \tag{6}$$

Intuitively, stabilizing the node with the highest cumulative sensitivity makes its encoding exact and effectively tightens all of these high sensitivity constraints. We can efficiently compute those scores without an additional backsubstitution pass.

Cost Adjusted Branching Any complete method will decide every property eventually. Hence its runtime is a key performance metric. Existing branching heuristics ignore this aspect and only consider the expected improvement in bound-tightness, but not the sometimes considerable differences in computational cost. We propose *Cost Adjusted Branching* (CAB), scaling the expected bound improvement with the inverse of the cost expected for the split, and then picking the branching decision yielding the highest expected bound improvement per cost. The true cost of a split consists of the direct cost of the next branching step and the change of cumulative cost for all consecutive steps. We find a local approximation considering just the former component, similar to only considering the one-step bound improvement, to be a good proxy. We approximate this direct cost by the number of floating-point operations required to compute the new bounds. However, any approximation of the expected gain and the expected cost can be used to instantiate CAB.

Enforcing splits Once the ReLU to split on has been determined, two subproblems are generated. One where a non-negative $\hat{z} \geq 0$ and one where non-positive $\hat{z} \leq 0$ pre-activation value have to be enforced. This is accomplished by setting the corresponding entries in the split matrix $\boldsymbol{S}$, used during the bounding process, to -1 and 1, respectively. As the intermediate bounds for all layers up to and including the one that was split remain unchanged, we do not recompute them. ´

3.3 SOUNDNESS AND COMPLETENESS

The soundness of the BaB approach follows directly from the soundness of the underlying bounding method discussed in Section 3.1. To show completeness, it is sufficient to consider the limit case where all ReLU nodes are split and the network becomes linear, making DEEPPOLY relaxations exact. However, to actually obtain exact bounds all split constraints have to be enforced, requiring an optimal choice of $\boldsymbol{\beta}$. This can be obtained as the fully split problem is convex and can hence be solved efficiently (Wang et al., 2021). Therefore, a property holds iff the exact bounds thus obtained are positive on all subproblems. We conclude that MN-BAB is a complete verifier.

Table 1: Natural accuracy [%] (Acc.), verified accuracy [%] (Ver.) and its empirical upper bound [%] and avg. runtime [s] of the first 1000 images of the test set.

Dataset	Model	Acc.	ϵ	ERAN		β-CROWN †		OVAL		MN-BAB (ours)			Upper Bound
				Ver.	Time	Ver.	Time	Ver.	Time	Branching	Ver.	Time	
MNIST	ConvSmall	98.0	0.120	**73.2**	38.4	71.6	46	69.8	26.2	BaBSR+CAB	70.4	**24.3**	73.2
	ConvBig	92.9	0.300	**78.6**	**6.0**	77.7	78	–	–	BaBSR+CAB	78.2	21.4	78.6
CIFAR10	ConvSmall	63.0	2/255	**47.2**	54.4	46.3	18	46.2	**7.2**	BaBSR+CAB	45.9	16.0	48.1
	ConvBig	63.1	2/255	48.2	128.1	50.3	55	50.6	42.0	ACS +CAB	**50.7**	**41.0**	55.0

† We report numbers from Wang et al. (2021).
– Assertion errors on some samples prevent reporting reliable numbers.

4 EXPERIMENTAL EVALUATION

We now present an extensive evaluation of our method. First, and perhaps surprisingly, we find that existing MILP based verification tools (Singh et al., 2019c; Müller et al., 2021) are more effective in verifying robustness on many established benchmark networks than what is considered state-of-the-art. This highlights that next-generation verifiers should focus on and be benchmarked using less regularized and more accurate networks. We take a step in this direction and propose such a network on which we analyze the effectiveness of the different components of MN-BAB. We show that the number of subproblems required for certification can be reduced by two orders of magnitude leveraging precise multi-neuron constraints, and then again 7-fold by our novel branching heuristic. Additionally, we show that using our cost-adjusted branching framework to make efficient branching decisions enables yet another speed up of 50%.

Experimental Setup MN-BAB is implemented in PyTorch (Paszke et al., 2019). We evaluate all benchmarks using a single NVIDIA RTX 2080Ti, 64 GB of memory, and 16 CPU cores. We attempt to falsify every property with an adversarial attack, before beginning certification. For every subproblem, we first lower-bound the objective using DEEPPOLY and then compute refined bounds using the method described in §3.

Benchmarks We benchmark MN-BAB on networks established in previous work (see Table 3 in App. A) on the MNIST (Deng, 2012) and CIFAR10 datasets (Krizhevsky et al., 2009).

We compare against β-CROWN (Wang et al., 2021), a BaB based state-of-the-art complete verifier, OVAL (Palma et al., 2021; De Palma et al., 2021; Bunel et al., 2020), a BaB framework based on a different class of multi-neuron constraints, and ERAN Singh et al. (2019c); Müller et al. (2021) combining the incomplete verifier PRIMA, whose tight multi-neuron constraints we leverage in our method, and a (partial) MILP encoding.

Importantly, we show how individual components of MN-BAB enable us to verify more challenging and accurate networks. To this end, we consider two new residual networks, `ResNet6-A`, and `ResNet6-B`, which have the same architecture but differ in regularization strength. Both networks were trained with adversarial training (Madry et al., 2017) using PGD, the GAMA loss (Sriramanan et al., 2020) and MixUp data augmentation (Zhang et al., 2021). `ResNet6-A` was trained using 8-steps and $\epsilon = 4/255$, whereas 20-steps and $\epsilon = 8/255$ were used for `ResNet6-B`.

Comparison on Existing Benchmarks In Table 1, we compare the verified accuracy and runtime of MN-BAB with that of state-of-the-art tools ERAN, β-CROWN, and OVAL. Perhaps surprisingly, we find that both MNIST and the smaller CIFAR10 benchmark network established over the last years (Singh et al., 2019b) can be verified completely or almost completely in less than 50s per sample using ERAN, making them less relevant as benchmarks for new verification tools. On these networks, all three BaB methods (including ours) are outperformed to a similar degree, with the reservation that we could not evaluate OVAL on MNIST `ConvBig` due to runtime errors. On the remaining unsolved network, where complete verification via a MILP encoding does not scale, MN-BAB obtains the highest certified accuracy and lowest runtime. Comparing the BaB methods, we observe an overall trend that more precise but also expensive underlying bounding methods are most effective on larger networks, where additional splits are less efficient, and complex inter-neuron interactions can be captured by the precise multi-neuron constraints.

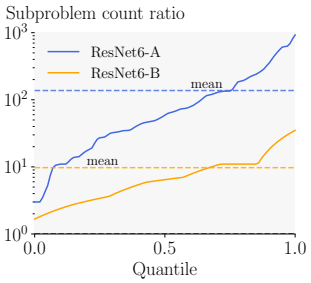

Figure 3: Ratio of subproblem required per property without and with MNC.

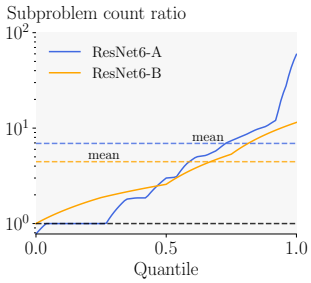

Figure 4: Ratio of subproblem required per property with BABSR and ACS.

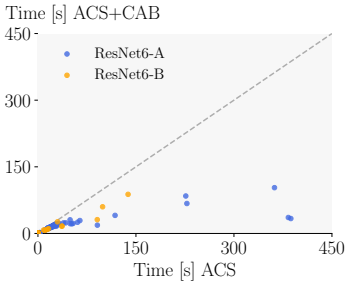

Figure 5: Effect of Cost Adjusted Branching on mean verification time with ACS.

Performance on Challenging Networks To measure the performance of verification tools in a more meaningful way (and illustrate that performance on highly regularized networks is not an ideal proxy), we consider the weakly and heavily regularized `ResNet6-A` and `ResNet6-B`, respectively, with identical architecture. Concretely, we show the effect different bounding procedures and branching approaches have in the two settings in Table 2. Since the publicly released code of both β-CROWN and OVAL do not support residual architectures without modifications, they could not be compared to. However, we note that BABSR without multi-neuron constraints is mathematically equivalent to β-CROWN. As verified accuracy and mean runtime are very coarse performance metrics, we also analyze the ratio of runtimes and number of subproblems required for verification on a per-property level, filtering out those where either method timed out or both methods verify before any branching occurred.

Table 2: Verified accuracy [%] (Ver.) and avg. runtime [s] on the first 100 images of the test set for $\epsilon = 2/255$).

Model	Acc.	Upper Bound	Branching Method	MNC	MN-BAB (ours)	
					Ver	Time
`ResNet6-A`	84	75	DEEPPOLY only	no	30	0.4
			BABSR	no	42	253.4
			BABSR +CAB	no	43	247.3
			BABSR	yes	45	225.0
			BABSR +CAB	yes	47	209.6
			ACS	yes	50	193.4
			ACS +CAB	yes	**53**	177.9
`ResNet6-B`	79	71	DEEPPOLY only	no	58	0.7
			BABSR	no	64	57.1
			BABSR +CAB	no	61	76.6
			BABSR	yes	64	57.5
			BABSR +CAB	yes	63	68.1
			ACS	yes	65	48.1
			ACS +CAB	yes	**66**	46.1

The trend of MN-BAB succeeding on more challenging networks becomes especially apparent here, where leveraging Multi-Neuron Constraints (MNCs) enables us to verify 26% more samples while being around 30% faster (see Table 2) on `ResNet6-A` while the differences on the more heavily regularized `ResNet6-B` are relatively small. When analyzing on a per-property level, shown in Fig. 6, the trend is even more obvious. For easy problems leveraging MNCs and ACS has little impact (points around the diagonal) however, it completely dominates on the harder properties where only using single-neuron constraints and BABSR takes up to 33-times longer (points below the diagonal).

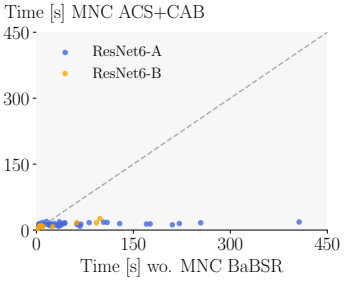

Figure 6: Per property verification times using MN-BAB over those without MNCs and using BABSR.

Effectiveness of Multi-Neuron Constraints In Fig. 3, we show the ratio between the number of subproblems required for certification with and without MNCs over the quantile of properties for `ResNet6-A` (blue) and `ResNet6-B` (orange). We observe that using MNCs reduces the number of subproblems for both networks. However, on `ResNet6-A` we achieve an average 137-fold reduction compared to a just 10-fold reduction on `ResNet6-B`. Despite the higher per bounding-step cost, MNCs increase the number of verified samples by up to 12% while reducing average certification times (see Table 2).

Effectiveness of ACS Branching In Fig. 4, we show the ratio between the number of subproblems considered when using BABSR vs. ACS over the quantile of properties. We observe that ACS yields significantly fewer subproblems on most (74%) or all properties on `ResNet6-A` and

ResNet6-B, respectively, leading to an on average 6.9 and 4.5-fold reduction and showing the effectiveness of our novel ACS branching heuristic. Average verification times are reduced by 12% and 19% on ResNet6-A and ResNet6-B, respectively. Note that the relatively small improvements in timings are due to timeouts for both methods yielding equally high runtimes which dominate the mean.

Effectiveness of Cost Adjusted Branching In Fig. 5, we show the per property verification times with ACS + CAB over those with ACS. Using CAB is faster (points below the dashed line) for all properties, sometimes significantly so, leading to an average speedup of around 50%. Analyzing the results in Table 2, we observe that CAB is particularly effective in combination with multi-neuron constraints, where bounding costs vary more significantly, and the more reliable ACS scores.

5 Related Work

Neural Network Verification Beyond the Single Neuron Convex Barrier After the so-called (Single Neuron) Convex Barrier has been described for incomplete relaxation-based methods by Salman et al. (2019), a range of approaches has been proposed that consider multiple neurons jointly to obtain tighter relaxations. Singh et al. (2019a); Müller et al. (2021) derive joint constraints over the input-output space of groups of neurons and refine their relaxation using the intersection of these constraints. Tjandraatmadja et al. (2020) merge the ReLU and preceding affine layer to consider multiple inputs but only one output at a time. While the two approaches are theoretically incomparable, the former yields empirically better results (Müller et al., 2021).

Early complete verification methods relied on off-the-shelf SMT (Katz et al., 2017; Ehlers, 2017) or MILP solvers (Dutta et al., 2018; Tjeng et al., 2017). However, these methods do not scale beyond small networks. In order to overcome these limitations, Bunel et al. (2020) formulated a BaB style framework for complete verification and showed it contains many previous methods as special cases. The basic idea is to recursively split the verification problem into easier subproblems on which cheap incomplete methods can show robustness. Since then, a range of partially (Xu et al., 2020b) or fully GPU-based (Wang et al., 2021; Palma et al., 2021) BaB frameworks have been proposed. The most closely related, Palma et al. (2021), leverages the multi-neuron constraints from Tjandraatmadja et al. (2020) but yields an optimization problem of different structure, as constraints only ever include single output neurons.

Branching Most ReLU branching methods proposed to date use the bound improvement of the two child subproblems as the metric to decide which node to branch on next. Full strong branching (Applegate et al., 1995) exhaustively evaluates this for all possible branching decisions. However, this is intractable for all but the smallest networks. Lu & Kumar (2019) train a GNN to imitate the behavior of full strong branching at a fraction of the cost, but transferability remains an open question and collecting training data to imitate is extremely costly. Bunel et al. (2020) proposed an efficiently computable heuristic score, locally approximating the bound improvement of a branching decision using the method of Wong & Kolter (2018). Henriksen & Lomuscio (2021) refine this approach by additionally approximating the indirect effect of the branching decision, however, this requires using two different bounding procedures. De Palma et al. (2021) introduced filtered-smart-branching (FSB), using BaBSR to selected branching candidates and then computing a more accurate heuristic score only for the selected candidates. Instead of targeting the largest bound improvement, Kouvaros & Lomuscio (2021) aim to minimize the number of unstable neurons by splitting the ReLU node with the most other ReLU nodes depending on it.

6 Conclusion

We propose the complete neural network verifier MN-BAB. Building on the Branch-and-Bound methodology, MN-BAB leverages tight multi-neuron constraints, a novel branching heuristic and an efficient dual solver able to utilize massively parallel hardware accelerators to enable the verification of particularly challenging networks. Our thorough empirical evaluation shows how MN-BAB is particularly effective in verifying challenging networks with high natural accuracy and practical relevance, reaching a new state-of-the-art in several settings.

7 ETHICS STATEMENT

Most machine learning based systems can be both employed with ethical as well as malicious purposes. Methods such as ours that enable the certification of robustness properties of neural networks are a step towards more safe and trustworthy AI systems and can hence amplify any such usage. Further, malicious actors might aim to convince regulators that the proposed approach is sufficient to show robustness to perturbations encountered during real world application, which could lead to insufficient regulation in safety critical domains.

8 REPRODUCIBILITY STATEMENT

We will make all code and trained networks required to reproduce our experiments available during the review process as supplementary material and provide instructions on how to run them. Upon publication, we will also release them publicly. We explain the basic experimental setup in Section 4 and provide more details in Section A. All datasets used in the experiments are publicly available. Random seeds are fixed where possible and provided in the supplementary material.

REFERENCES

David Applegate, Robert Bixby, Vašek Chvátal, and William Cook. Finding cuts in the tsp (a preliminary report). Technical report, Citeseer, 1995.

Rudy Bunel, P Mudigonda, Ilker Turkaslan, P Torr, Jingyue Lu, and Pushmeet Kohli. Branch and bound for piecewise linear neural network verification. *Journal of Machine Learning Research*, 21(2020), 2020.

Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1310–1320. PMLR, 2019. URL http://proceedings.mlr.press/v97/cohen19c.html.

Sumanth Dathathri, Krishnamurthy Dvijotham, Alexey Kurakin, Aditi Raghunathan, Jonathan Uesato, Rudy Bunel, Shreya Shankar, Jacob Steinhardt, Ian J. Goodfellow, Percy Liang, and Pushmeet Kohli. Enabling certification of verification-agnostic networks via memory-efficient semidefinite programming. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/397d6b4c83c91021fe928a8c4220386b-Abstract.html.

Alessandro De Palma, Rudy Bunel, Alban Desmaison, Krishnamurthy Dvijotham, Pushmeet Kohli, Philip HS Torr, and M Pawan Kumar. Improved branch and bound for neural network verification via lagrangian decomposition. *arXiv preprint arXiv:2104.06718*, 2021.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Souradeep Dutta, Susmit Jha, Sriram Sankaranarayanan, and Ashish Tiwari. Output range analysis for deep feedforward neural networks. In *NASA Formal Methods Symposium*, pp. 121–138. Springer, 2018.

Ruediger Ehlers. Formal verification of piece-wise linear feed-forward neural networks. In *International Symposium on Automated Technology for Verification and Analysis*, pp. 269–286. Springer, 2017.

Timon Gehr, Matthew Mirman, Dana Drachsler-Cohen, Petar Tsankov, Swarat Chaudhuri, and Martin Vechev. 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.

Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. On the effectiveness of interval bound propagation for training verifiably robust models. *arXiv preprint arXiv:1810.12715*, 2018.

Patrick Henriksen and Alessio Lomuscio. Deepsplit: An efficient splitting method for neural network verification via indirect effect analysis. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI21), To appear*, 2021.

Guy Katz, Clark Barrett, David L Dill, Kyle Julian, and Mykel J Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks. In *International Conference on Computer Aided Verification*, pp. 97–117. Springer, 2017.

Panagiotis Kouvaros and Alessio Lomuscio. Towards scalable complete verification of relu neural networks via dependency-based branching. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI21), To Appear*, 2021.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

A. H. Land and A. G. Doig. An automatic method of solving discrete programming problems. *Econometrica*, 28(3):497–520, 1960. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1910129.

Jingyue Lu and M Pawan Kumar. Neural network branching for neural network verification. *arXiv preprint arXiv:1912.01329*, 2019.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Mark Niklas Müller, Gleb Makarchuk, Gagandeep Singh, Markus Püschel, and Martin Vechev. Prima: Precise and general neural network certification via multi-neuron convex relaxations. *arXiv preprint arXiv:2103.03638*, 2021.

Alessandro De Palma, Harkirat S. Behl, Rudy R. Bunel, Philip H. S. Torr, and M. Pawan Kumar. Scaling the convex barrier with active sets. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=uQfOy7LrlTR.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Hadi Salman, Greg Yang, Huan Zhang, Cho-Jui Hsieh, and Pengchuan Zhang. A convex relaxation barrier to tight robustness verification of neural networks. *arXiv preprint arXiv:1902.08722*, 2019.

Gagandeep Singh, Timon Gehr, Matthew Mirman, Markus Püschel, and Martin T Vechev. Fast and effective robustness certification. *NeurIPS*, 1(4):6, 2018.

Gagandeep Singh, Rupanshu Ganvir, Markus Püschel, and Martin Vechev. Beyond the single neuron convex barrier for neural network certification. 2019a.

Gagandeep Singh, Timon Gehr, Markus Püschel, and Martin Vechev. An abstract domain for certifying neural networks. *Proceedings of the ACM on Programming Languages*, 3(POPL):1–30, 2019b.

Gagandeep Singh, Timon Gehr, Markus Püschel, and Martin Vechev. Boosting robustness certification of neural networks. In *International Conference on Learning Representations (ICLR)*. 2019c.

Gaurang Sriramanan, Sravanti Addepalli, Arya Baburaj, and Venkatesh Babu R. Guided adversarial attack for evaluating and enhancing adversarial defenses. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/ea3ed20b6b101a09085ef09c97da1597-Abstract.html.

Yusuke Tashiro, Yang Song, and Stefano Ermon. Output diversified initialization for adversarial attacks. *ArXiv preprint*, abs/2003.06878, 2020.

Christian Tjandraatmadja, Ross Anderson, Joey Huchette, Will Ma, Krunal Patel, and Juan Pablo Vielma. The convex relaxation barrier, revisited: Tightened single-neuron relaxations for neural network verification. *arXiv preprint arXiv:2006.14076*, 2020.

Vincent Tjeng, Kai Xiao, and Russ Tedrake. Evaluating robustness of neural networks with mixed integer programming. *arXiv preprint arXiv:1711.07356*, 2017.

Shiqi Wang, Huan Zhang, Kaidi Xu, Xue Lin, Suman Jana, Cho-Jui Hsieh, and J Zico Kolter. Beta-crown: Efficient bound propagation with per-neuron split constraints for neural network robustness verification. In *ICML 2021 Workshop on Adversarial Machine Learning*, 2021.

Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, pp. 5286–5295. PMLR, 2018.

Kaidi Xu, Zhouxing Shi, Huan Zhang, Yihan Wang, Kai-Wei Chang, Minlie Huang, Bhavya Kailkhura, Xue Lin, and Cho-Jui Hsieh. Automatic perturbation analysis for scalable certified robustness and beyond. *Advances in Neural Information Processing Systems*, 33, 2020a.

Kaidi Xu, Huan Zhang, Shiqi Wang, Yihan Wang, Suman Jana, Xue Lin, and Cho-Jui Hsieh. Fast and complete: Enabling complete neural network verification with rapid and massively parallel incomplete verifiers. *arXiv preprint arXiv:2011.13824*, 2020b.

Linjun Zhang, Zhun Deng, Kenji Kawaguchi, Amirata Ghorbani, and James Zou. How does mixup help with robustness and generalization? In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=8yKEo06dKNo.

Table 3: Overview of the experimental configuration for every network.

Dataset	Model	Training	Timeout	Batch sizes	#Activation Layers
MNIST	`ConvSmall`	DiffAI	360	[128, 256, 512]	3
	`ConvBig`	PGD	500	[1, 2, 2, 3, 6, 12]	6
CIFAR10	`ConvSmall`	PGD	360	[100, 200, 400]	3
	`ConvBig`	PGD	2000	[3, 3, 6, 6, 8, 16]	6
	`ResNet6-A`	PGD	600	[4, 4, 8, 8, 16, 128]	6
	`ResNet6-B`	PGD	600	[4, 4, 8, 8, 16, 128]	6

A Experiment Details

In Table 3 we show the per-sample timeout and the batch sizes that were used in the experiments. The timeouts for the first 4 networks were chosen to approximately match the average runtimes reported by Wang et al. (2021) facilitate comparability.

Since we keep intermediate bounds of neurons before the split layer fixed, as described in Section 3.2, the memory requirements for splitting at different layers can vary. We exploit this fact and choose batch sizes for our bounding procedure depending on the layer where the split occurred.

In order to falsify properties more quickly, we run a strong adversarial attack with the following parameters before attempting certification: We apply two targeted version (towards all classes) of PGD (Madry et al., 2017) using margin loss (Gowal et al., 2018) and GAMA loss (Sriramanan et al., 2020), both with 5 restarts, 50 steps, and 10 step output diversification (Tashiro et al., 2020).

A.1 Architectures

In this section, we provide an overview over all the architectures evaluated in Section 4. The architecture of the convolutional networks for MNIST and CIFAR10 are detailed in Table 4. And the architecture of both `ResNet6-A` and `ResNet6-B` is given in Table 5

Table 4: Network architectures of the convolutional networks for CIFAR10, MNIST. All layers listed below are followed by an activation layer. The output layer is omitted. 'Conv c h×w/s/p' corresponds to a 2D convolution with c output channels, an h×w kernel size, a stride of s in both dimensions, and an all around zero padding of p.

`ConvSmall`	`ConvBig`
Conv 16 4×4/2/0	Conv 32 3×3/1/1
Conv 32 4×4/2/0	Conv 32 4×4/2/1
FC 100	Conv 64 3×3/1/1
	Conv 64 4×4/2/1
	FC 512
	FC 512

Table 5: Network architecture of the `ResNet6`. All layers listed below are followed by a ReLU activation layer, except if they are followed by a ResAdd layer. The output layer is omitted. 'Conv c h×w/s/p' corresponds to a 2D convolution with c output channels, an h×w kernel size, a stride of s in both dimensions, and an all around zero padding of p.

`ResNet6`	
Conv 16 3×3/1/1	
Conv 32 1×1/2/0	Conv 32 3×3/2/1
	Conv 32 3×3/1/1
ResAdd	
Conv 64 1×1/2/0	Conv 64 3×3/2/1
	Conv 64 3×3/1/1
ResAdd	
FC 100	