Piecewise Linear Neural Network Verification
A Comparative Study

Rudy Bunel, Ilker Turkaslan, Philip HS Torr, Pushmeet Kohli, M Pawan Kumar
Problem

Assume that you are given:

- The weights of a Neural Network
- Some property that should hold over a certain region of the input domain

You want to prove that the property always holds.

Example applications:

- If I’m less than $d$ away from a point, can I guarantee the same prediction?
- If inputs are in range, can I guarantee the predicted class is $c$?
- Is the numerical value of output $o$ always below limit value $l$?
Formalism

For a Neural Network that implements the function

\[ y = f(x) \]

We want to prove that:

\[ x \in C \]
\[ y = f(x) \]

\[ \Rightarrow \]

\[ P(y) \]
Example – Robustness to Adversarial Examples

\[ \| x - 1 \|_\infty \leq \epsilon \]

\[ x \in C \]
\[ y = f(x) \]
\[ \Rightarrow \]
\[ P(y) \]

\[ \begin{pmatrix} 1 \\ -1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \geq 0 \]
Example – Bound on output

Prove that $y > -5$
What about more complex properties?

OR clauses:

\[ P(y) \triangleq [y_1 \geq 0] \lor [y_2 \geq 0] \lor [y_3 \geq 0] \]

\[ \iff \]

\[ \max(y_1, y_2, y_3) \geq 0 \]

\[ \rightarrow \] Can be implemented by adding Maxpooling layer at the end
What about more complex properties?

AND clauses:

\[
P(y) \triangleq [y_1 \geq 0] \land [y_2 \geq 0] \land [y_3 \geq 0]
\]

\[\iff\]

\[
\min(y_1, y_2, y_3) \geq 0
\]

\[\iff\]

\[
-\max(-y_1, -y_2, -y_3) \geq 0
\]

Can be implemented by adding Linear + Maxpooling + Linear
Canonical problem

For any property that is a Boolean formula over linear inequalities,

→ We can just build an equivalent network for which we prove a bound
Example – Bound on output

\[ x \in C \]
\[ y = f(x) \]
\[ \Rightarrow \]
\[ P(y) \]

Prove that \( y > -5 \)
Strategy: counterexample search

Create a feasibility problem for the existence of a counterexample.

If the problem is feasible:
    then a counterexample is shown, the property is False

If the problem is infeasible:
    then no counterexample can exist, the property is True
Counter example search

Prove that $y > -5$

$-2 \leq x_1 \leq 2$
$-2 \leq x_2 \leq 2$

$a_{in} = x_1 + x_2$

$b_{in} = x_1 - x_2$

$a_{out} = \max(a_{in}, 0)$

$b_{out} = \max(b_{in}, 0)$

$y = -a_{out} - b_{out}$

$y \leq -5$
Difficulty

\[-2 \leq x_1 \leq 2\]
\[-2 \leq x_2 \leq 2\]
\[a_{in} = x_1 + x_2\]
\[b_{in} = x_1 - x_2\]
\[a_{out} = \max(a_{in}, 0)\]
\[b_{out} = \max(b_{in}, 0)\]
\[y = -a_{out} - b_{out}\]
\[y \leq -5\]

\[a_{out} = \max(a_{in}, 0)\]
\[b_{out} = \max(b_{in}, 0)\]

The problem is in the non-linearities!
Strategy 1: MIP solver

In Gurobi we trust

Chih-Hong Cheng, Georg Nuhrenberg, and Harald Ruess. *Maximum resilience of artificial neural networks.* Automated Technology for Verification and Analysis, 2017.

Alessio Lomuscio and Lalit Maganti. *An approach to reachability analysis for feed-forward relu neural networks.* arXiv:1706.07351, 2017.
Strategy 1: Mixed integer encoding

\[ a_{out} = \max(a_{in}, 0) \]

\[ a_{out} \geq a_{in} \]
\[ a_{out} \geq 0 \]

\[ a_{out} \leq a_{in} + (1 - \delta_a)M_a \]
\[ a_{out} \leq \delta_a M_a \]

\[ \delta_a \in \{0, 1\} \]

where \( M_a \) is big enough
Strategy 1: Mixed integer encoding

If $\delta_a = 0$

$$0 \leq a_{out} \leq 0$$

If $\delta_a = 1$

$$a_{in} \leq a_{out} \leq a_{in}$$

$$a_{out} \geq a_{in}$$

$$a_{out} \geq 0$$

$$a_{out} \leq a_{in} + (1 - \delta_a)M_a$$

$$a_{out} \leq \delta_a M_a$$

$$\delta_a \in \{0, 1\}$$

Problem: Mixed Integer Programming is also hard.
Strategy 2: Reluplex

Modify assignments until all constraints are satisfied.
If you feel stuck, divide the problem into two easier ones

Guy Katz, Clark Barrett, David Dill, Kyle Julian, and Mykel Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks. Computer Aided Verification, 2017.
Strategy 2: Reluplex

| $x_1$ | $x_2$ | $a_{in}$ | $a_{out}$ | $b_{in}$ | $b_{out}$ | $y$ |
|-------|-------|----------|-----------|----------|-----------|----|
| 0     | 0     | 0        | 0         | 0        | 0         | 0  |

Fix Linear constraints

- $-2 \leq x_1 \leq 2$
- $-2 \leq x_2 \leq 2$
- $a_{in} = x_1 + x_2$
- $b_{in} = x_1 - x_2$
- $a_{out} = \max(a_{in}, 0)$
- $b_{out} = \max(b_{in}, 0)$
- $y = -a_{out} - b_{out}$
- $y \leq -5$
Strategy 2: Reluplex

| $x_1$ | $x_2$ | $a_{in}$ | $a_{out}$ | $b_{in}$ | $b_{out}$ | $y$ |
|-------|-------|---------|-----------|---------|----------|-----|
| 0     | 0     | 0       | 1         | 0       | 4        | -5  |

Fix one non-linear constraint (b)

| $x_1$ | $x_2$ | $a_{in}$ | $a_{out}$ | $b_{in}$ | $b_{out}$ | $y$ |
|-------|-------|---------|-----------|---------|----------|-----|
| 0     | 0     | 0       | 1         | 0       | 4        | -5  |

$-2 \leq x_1 \leq 2$

$-2 \leq x_2 \leq 2$

$a_{in} = x_1 + x_2$

$b_{in} = x_1 - x_2$

$a_{out} = \max(a_{in}, 0)$

$b_{out} = \max(b_{in}, 0)$

$y = -a_{out} - b_{out}$

$y \leq -5$
Strategy 2: Reluplex

|   |   | $a_{in}$ | $a_{out}$ | $b_{in}$ | $b_{out}$ | $y$ |
|---|---|---------|---------|---------|---------|-----|
| 0 | 0 | 0       | 1       | 4       | 4       | -5  |
| 2 | -2| 0       | 1       | 4       | 4       | -5  |

Fix Linear constraints

$$-2 \leq x_1 \leq 2$$
$$-2 \leq x_2 \leq 2$$

$$a_{in} = x_1 + x_2$$
$$b_{in} = x_1 - x_2$$

$$a_{out} = \max(a_{in}, 0)$$
$$b_{out} = \max(b_{in}, 0)$$

$$y = -a_{out} - b_{out}$$
$$y \leq -5$$
Strategy 2: Reluplex

If no progress is made (the same non-linearity is fixed too many times), the problem is split into two and the search continues.

\[
\begin{align*}
-2 \leq x_1 &\leq 2 \\
-2 \leq x_2 &\leq 2 \\
a_{in} &= x_1 + x_2 \\
b_{in} &= x_1 - x_2 \\
a_{out} &= \max(a_{in}, 0) \\
b_{out} &= \max(b_{in}, 0) \\
y &= -a_{out} - b_{out} \\
y &\leq -5
\end{align*}
\]
Strategy 3: Planet

Divide the problem until you reach infeasibilities, then backtrack

Ruediger Ehlers.
*Formal verification of piece-wise linear feed-forward neural networks.*
Automated Technology for Verification and Analysis, 2017
Strategy 3: Planet

Building a linear approximation of the network:

→ Replace non-linearities by set of constraints

Original ReLU Non-linearity

Non Linearity approximation
Strategy 3: Planet

\[-2 \leq x_1 \leq 2\]
\[-2 \leq x_2 \leq 2\]
\[a_{in} = x_1 + x_2\]
\[b_{in} = x_1 - x_2\]
\[a_{out} = \max(a_{in}, 0)\]
\[b_{out} = \max(b_{in}, 0)\]
\[y = -a_{out} - b_{out}\]
\[y \leq -5\]

Network Linear Approximation

\[-2 \leq x_1 \leq 2\]
\[-2 \leq x_2 \leq 2\]
\[a_{in} = x_1 + x_2\]
\[b_{in} = x_1 - x_2\]
\[a_{out} \geq 0, \quad a_{out} \geq a_{in}\]
\[a_{out} \leq \frac{a_{in}}{2} + 2\]
\[b_{out} \geq 0, \quad b_{out} \geq b_{in}\]
\[b_{out} \leq \frac{b_{in}}{2} + 2\]
\[y = -a_{out} - b_{out}\]
\[y \leq -5\]
Strategy 3: Planet

\[
\begin{align*}
-2 &\leq x_1 \leq 2 \\
-2 &\leq x_2 \leq 2 \\
a_{\text{in}} &= x_1 + x_2 \\
b_{\text{in}} &= x_1 - x_2 \\
a_{\text{out}} &\geq 0, \quad a_{\text{out}} \geq a_{\text{in}} \\
a_{\text{out}} &\leq \frac{a_{\text{in}}}{2} + 2 \\
b_{\text{out}} &\geq 0, \quad b_{\text{out}} \geq b_{\text{in}} \\
b_{\text{out}} &\leq \frac{b_{\text{in}}}{2} + 2 \\
y &= -a_{\text{out}} - b_{\text{out}} \\
y &\leq -5
\end{align*}
\]
Strategy 3: Planet

Is it feasible?
No

\[-2 \leq x_1 \leq 2\]
\[-2 \leq x_2 \leq 2\]
\[a_{in} = x_1 + x_2\]
\[b_{in} = x_1 - x_2\]
\[a_{out} \geq 0, \quad a_{out} \geq a_{in}\]
\[a_{out} \leq \frac{a_{in}}{2} + 2\]
\[b_{out} \geq 0, \quad b_{out} \geq b_{in}\]
\[b_{out} \leq \frac{b_{in}}{2} + 2\]
\[y = -a_{out} - b_{out}\]
\[y \leq -5\]
\[a_{out} = 0\]
Strategy 3: Planet

Is it feasible?

Yes \( (x_1 = 2, x_2 = 2, a_{\text{out}} = 4, b_{\text{out}} = 2) \)

\[-2 \leq x_1 \leq 2\]
\[-2 \leq x_2 \leq 2\]
\[a_{\text{in}} = x_1 + x_2\]
\[b_{\text{in}} = x_1 - x_2\]
\[a_{\text{out}} \geq 0, \quad a_{\text{out}} \geq a_{\text{in}}\]
\[a_{\text{out}} \leq \frac{a_{\text{in}}}{2} + 2\]
\[b_{\text{out}} \geq 0, \quad b_{\text{out}} \geq b_{\text{in}}\]
\[b_{\text{out}} \leq \frac{b_{\text{in}}}{2} + 2\]
\[y = -a_{\text{out}} - b_{\text{out}}\]
\[y \leq -5\]
\[a_{\text{out}} = a_{\text{in}}\]
Strategy 3: Planet

Is it feasible? No

\[-2 \leq x_1 \leq 2\]
\[-2 \leq x_2 \leq 2\]
\[a_{in} = x_1 + x_2\]
\[b_{in} = -x_1 - x_2\]
\[a_{out} \geq 0, \quad a_{out} \geq a_{in}\]
\[a_{out} \leq \frac{a_{in}}{2} + 2\]
\[b_{out} \geq 0, \quad b_{out} \geq b_{in}\]
\[b_{out} \leq \frac{b_{in}}{2} + 2\]
\[y = -a_{out} - b_{out}\]
\[y \leq -5\]
\[a_{out} = a_{in}\]
\[b_{out} = 0\]
Strategy 3: Planet

-2 ≤ x₁ ≤ 2
-2 ≤ x₂ ≤ 2

a_{in} = x₁ + x₂
b_{in} = -x₁ - x₂

a_{out} ≥ 0, a_{out} ≥ a_{in}
a_{out} ≤ \frac{a_{in}}{2} + 2
b_{out} ≥ 0, b_{out} ≥ b_{in}
b_{out} ≤ \frac{b_{in}}{2} + 2

y = -a_{out} - b_{out}
y ≤ -5

a_{out} = a_{in}
b_{out} = b_{in}

Is it feasible?
No
Strategy 3: Planet

There is no feasible assignment
No counterexample can exist.
The property is True

\[-2 \leq x_1 \leq 2\]
\[-2 \leq x_2 \leq 2\]
\[a_{in} = x_1 + x_2\]
\[b_{in} = -x_1 - x_2\]
\[a_{out} \geq 0, \quad a_{out} \geq a_{in}\]
\[a_{out} \leq \frac{a_{in}}{2} + 2\]
\[b_{out} \geq 0, \quad b_{out} \geq b_{in}\]
\[b_{out} \leq \frac{b_{in}}{2} + 2\]
\[y = -a_{out} - b_{out}\]
\[y \leq -5\]
Strategy 4: Branch and Bound

Evaluate worst case
If it’s too bad, split the problem into smaller ones
Strategy 4: Branch and Bound

Use Planet approximation, and compute the minimum of $y$

$$-2 \leq x_1 \leq 2$$
$$-2 \leq x_2 \leq 2$$
$$a_{in} = x_1 + x_2$$
$$b_{in} = x_1 - x_2$$
$$a_{out} \geq 0, \quad a_{out} \geq a_{in}$$
$$a_{out} \leq \frac{a_{in}}{2} + 2$$
$$b_{out} \geq 0, \quad b_{out} \geq b_{in}$$
$$b_{out} \leq \frac{b_{in}}{2} + 2$$
$$y = -a_{out} - b_{out}$$

Minimum of $y$, subject to all the constraints:

$$y_{min} = -6$$

We look for $y \leq -5$

So there might be counterexample in the region
Strategy 4: Branch and Bound

Use Planet approximation, and compute the minimum of $y$

$$-2 \leq x_1 \leq 0$$
$$-2 \leq x_2 \leq 2$$
$$a_{in} = x_1 + x_2$$
$$b_{in} = -x_1 - x_2$$
$$a_{out} \geq 0, \quad a_{out} \geq a_{in}$$
$$a_{out} \leq \frac{a_{in}}{3} + \frac{4}{3}$$
$$b_{out} \geq 0, \quad b_{out} \geq b_{in}$$
$$b_{out} \leq \frac{b_{in}}{3} + \frac{4}{3}$$
$$y = -a_{out} - b_{out}$$

Minimum of $y$, subject to all the constraints:

$$y_{min} = -2.66$$

We look for $y \leq -5$

So no possible counterexample on this region.
Strategy 4: Branch and Bound

Use Planet approximation, and compute the minimum of $y$

\[
0 \leq x_1 \leq 2 \\
-2 \leq x_2 \leq 2 \\
a_{in} = x_1 + x_2 \\
b_{in} = -x_1 - x_2 \\
a_{out} \geq 0, \ a_{out} \geq a_{in} \\
a_{out} \leq \frac{2}{3} a_{in} + \frac{4}{3} \\
b_{out} \geq 0, \ b_{out} \geq b_{in} \\
b_{out} \leq \frac{2}{3} b_{in} + \frac{4}{3} \\
y = -a_{out} - b_{out}
\]

Minimum of $y$, subject to all the constraints:

\[y_{min} = -5.33\]

We look for $y \leq -5$
So no possible counterexample on this region
Strategy 4: Branch and Bound

Use Planet approximation, and compute the minimum of $y$

$$0 \leq x_1 \leq 2$$
$$-2 \leq x_2 \leq 0$$
$$a_{in} = x_1 + x_2$$
$$b_{in} = -x_1 - x_2$$
$$a_{out} \geq 0, \ a_{out} \geq a_{in}$$
$$a_{out} \leq \frac{a_{in}}{2} + 1$$
$$b_{out} \geq 0, \ b_{out} \geq b_{in}$$
$$b_{out} \leq b_{in}$$
$$y = -a_{out} - b_{out}$$

Minimum of $y$, subject to all the constraints:
Etc ...
Strategy 4: Branch and Bound

Should the approximation be rebuilt at each step?

Much faster to reuse the existing approximation

But the bounds don’t really improve
Experimental Showdown

Who’s the fastest of them all?
It depends on the properties \(\_\_\_\(\ツ\)_/\_\)

### ACAS dataset

| Data Set | Success Rate | Average runtime SAT | Average runtime UNSAT | Number of Wins SAT | Number of Wins UNSAT |
|----------|--------------|---------------------|-----------------------|--------------------|----------------------|
| Reluplex | 79.79%       | 2077.7 s            | 1868.3 s              | 3                  | 6                    |
| Planet   | 46.28%       | 7200 s              | 2532.8 s              | 0                  | 1                    |
| MIP      | 49.47%       | 6014.0 s            | 2708.2 s              | 3                  | 2                    |
| BaB      | 83.51%       | 128.14 s            | 733.55 s              | 29                 | 118                  |

### Collision Detection

| Data Set | Success Rate | Average runtime SAT | Average runtime UNSAT | Number of Wins SAT | Number of Wins UNSAT |
|----------|--------------|---------------------|-----------------------|--------------------|----------------------|
| Reluplex | 99.4 %       | 1.14 s              | 1.17 s                | 70                 | 30                   |
| Planet   | 100 %        | 0.50 s              | 0.18 s                | 79                 | 275                  |
| MIP      | 100 %        | 0.66 s              | 0.65 s                | 17                 | 11                   |
| BaB      | 100 %        | 5.27 s              | 29.05 s               | 0                  | 0                    |

### TwinStream

| Data Set | Success Rate | Average Runtime | Number of Wins |
|----------|--------------|-----------------|----------------|
| Reluplex*| 53.09%       | 2061.8s         | 1              |
| Planet   | 62.96%       | 2878.5s         | 13             |
| MIP      | 65.43%       | 2647.0s         | 41             |
| BaB      | 28.40%       | 5248.1s         | 4              |
What next?

Branch-and-Bound is a flexible framework:

• How to choose the splitting to prioritize?

• Can we estimate a good scale to start the bounding process?

• Can we get better lower bounds?
Better bounds (in progress)

What does better bounds means?

Zico Kolter, Eric Wong.  
*Provable defenses against adversarial examples via the convex outer adversarial polytope.*  
Arxiv:1711.00851
Better Bounds

\[ \text{min } y \]

such that

\[ -2 \leq x_1 \leq 2 \]
\[ -2 \leq x_2 \leq 2 \]
\[ a_{in} = x_1 + x_2 \]
\[ b_{in} = x_1 - x_2 \]
\[ a_{out} \geq 0, \quad a_{out} \geq a_{in} \]
\[ a_{out} \leq \frac{a_{in}}{2} + 2 \]
\[ b_{out} \geq 0, \quad b_{out} \geq b_{in} \]
\[ b_{out} \leq \frac{b_{in}}{2} + 2 \]
\[ y = -a_{out} - b_{out} \]

Primal problem

\[ \text{max } -2(\mu_{0,b} + \mu_{0,a}) - 2(\epsilon_{u1} + \epsilon_{u2}) - 2(\epsilon_{l1} + \epsilon_{l2}) \]

such that

\[ \mu_{0,a} = 1 + \rho_{0,a} + \tau_{0,a} \]
\[ \mu_{0,b} = 1 + \rho_{0,b} + \tau_{0,b} \]
\[ \mu_{0,a} = 2 \left( \tau_{0,a} + \lambda_{0,a} \right) \]
\[ \mu_{0,b} = 2 \left( \tau_{0,b} + \lambda_{0,b} \right) \]
\[ \lambda_{0,a} + \lambda_{0,b} = \epsilon_{u1} - \epsilon_{l1} \]
\[ \lambda_{0,a} - \lambda_{0,b} = \epsilon_{u2} - \epsilon_{l2} \]
\[ \epsilon \geq 0, \quad \mu \geq 0, \quad \rho \geq 0, \quad \tau \geq 0 \]

Dual problem
Better Bounds

We know that optimal of primal problem is the same as the optimal of the dual problem.

(Strong Duality)

But we also know that any feasible solution to the dual problem is a valid lower bound on the optimal solution of the primal problem

(Weak Duality)
Better Bounds

Global minimum of the Neural Network $\geq$ Minimum (unique) of the NN linear approximation $=$ Maximum (unique) of the dual of the NN linear approximation $\geq$ Any feasible point of the dual of the NN linear approximation

Don’t need to solve the dual problem exactly! Any heuristic will work!

How well does it perform?
Approximate Dual solution for BaB

The bound computed are worse

But they are significantly faster to compute!
Collaborators

Pushmeet Kohli  
*Deepmind*

Philip Torr  
*University of Oxford*

Pawan Kumar  
*University of Oxford*  
*Alan Turing Institute*

Ilker Turkaslan  
*University of Oxford*
Thank you for your attention. Any questions?

rudy@robots.ox.ac.uk