EFFICIENT FORMAL SAFETY ANALYSIS OF NEURAL NETWORKS

Shiqi Wang, Kexin Pei, Justin Whitehouse, Junfeng Yang, Suman Jana

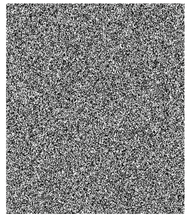
COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

ADVERSARIAL EXAMPLES:

Existing violations of NN safety

Dog

\+



L Bounded Noise

=

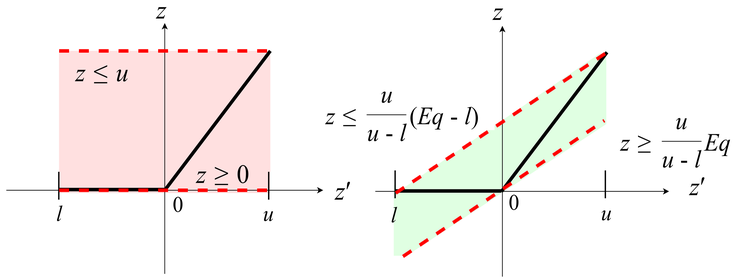
Sushi

How to formally guarantee the absence of violations within bounded input ranges?

Output nonconvexity => need tight approximation

1. SYMBOLIC LINEAR RELAXATION

Tighter interval analysis for ReLU propagations
Partial input dependencies are preserved
Used to identify crucial **overestimated nodes**
Overestimated nodes: the node performs nonlinearity



(a) Naive concretizaion

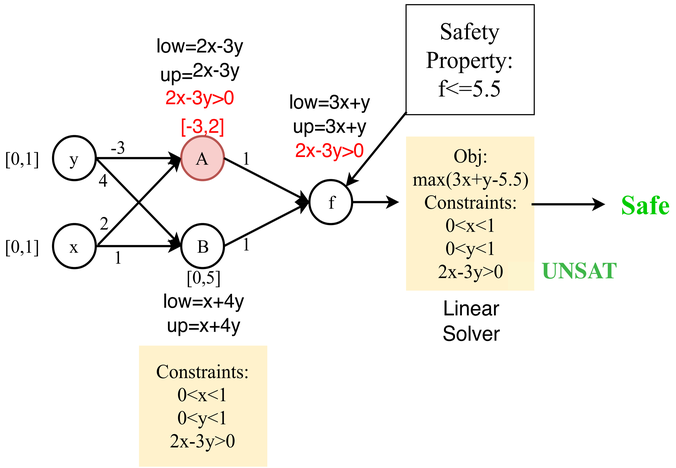
Dependencies ignored

(b) Symbolic linear relaxation

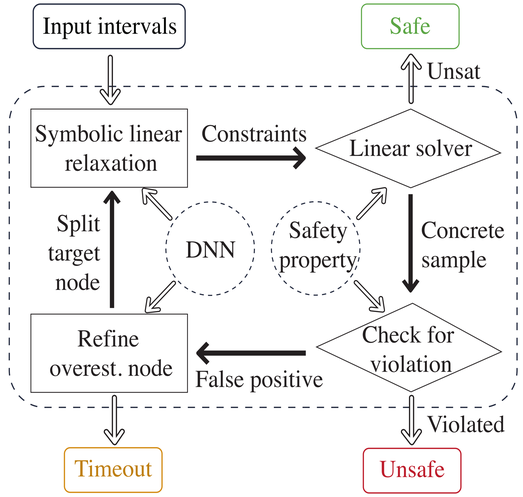
Partial dependencies

2. DIRECTED CONSTRAINT RELAXATION

Locate influential overestimated nodes
Split each nonlinear ReLU into two linear cases
Solve each case with **linear solver**



HOW NEURIFY SOLVES THIS PROBLEM?



Interval & Linear solver

Given: (1) Input ranges
(2) Targeted network and
(3) predefined safety property

Neurify: (1) Locate overestimated nodes with symolic intervals and
(2) Iteratively refine approximated output ranges with linear solver

Terminate: (1) Proved safe
(2) Proved unsafe with counter-examples and (3) Timeout

RESULTS

ACAS Xu: **5000** times faster than Reluplex and **20** times faster than ReluVal

DAVE: First system to scale to network over **10,000** ReLUs. Various safety properties (e.g., L_1, L_∞, lightening, contrast) can be formally analyzed.

Code availabe at https://github.com/tcwangshiqi-columbia/Neurify