Optimization + Abstraction: A Synergistic Approach for Analyzing Neural Network Robustness

Greg Anderson[1], Shankara Pailoor[1], Isil Dillig[1], Swarat Chaudhuri[2]

Overview

The goal: Prove that a neural network is *robust* in some region, i.e., all the points in that region are put into the same class by the network.

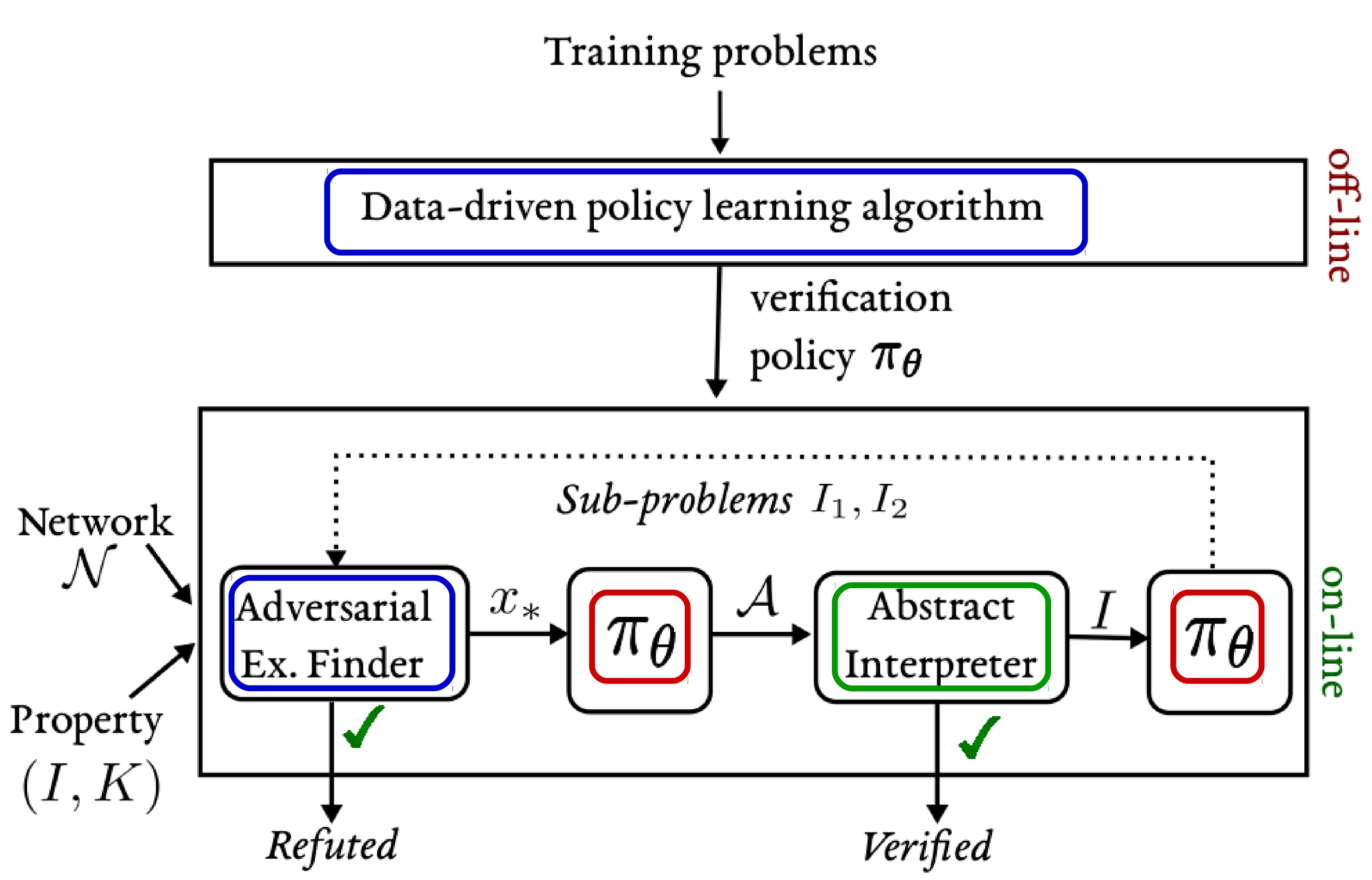
Key Idea: Combine Optimization and Abstraction to verify robustness

Two kinds of optimization
- Searching for counterexamples can efficiently falsify properties
- Learning a verification policy

Abstract interpretation can efficiently prove properties, but can't falsify them.

What if we can neither find a counterexample nor prove the property?
- Split the region into two pieces and verify each independently
- The network is robust in the original region iff it is robust in each subregion.

The verification policy tells us how to analyze each region and where to split.

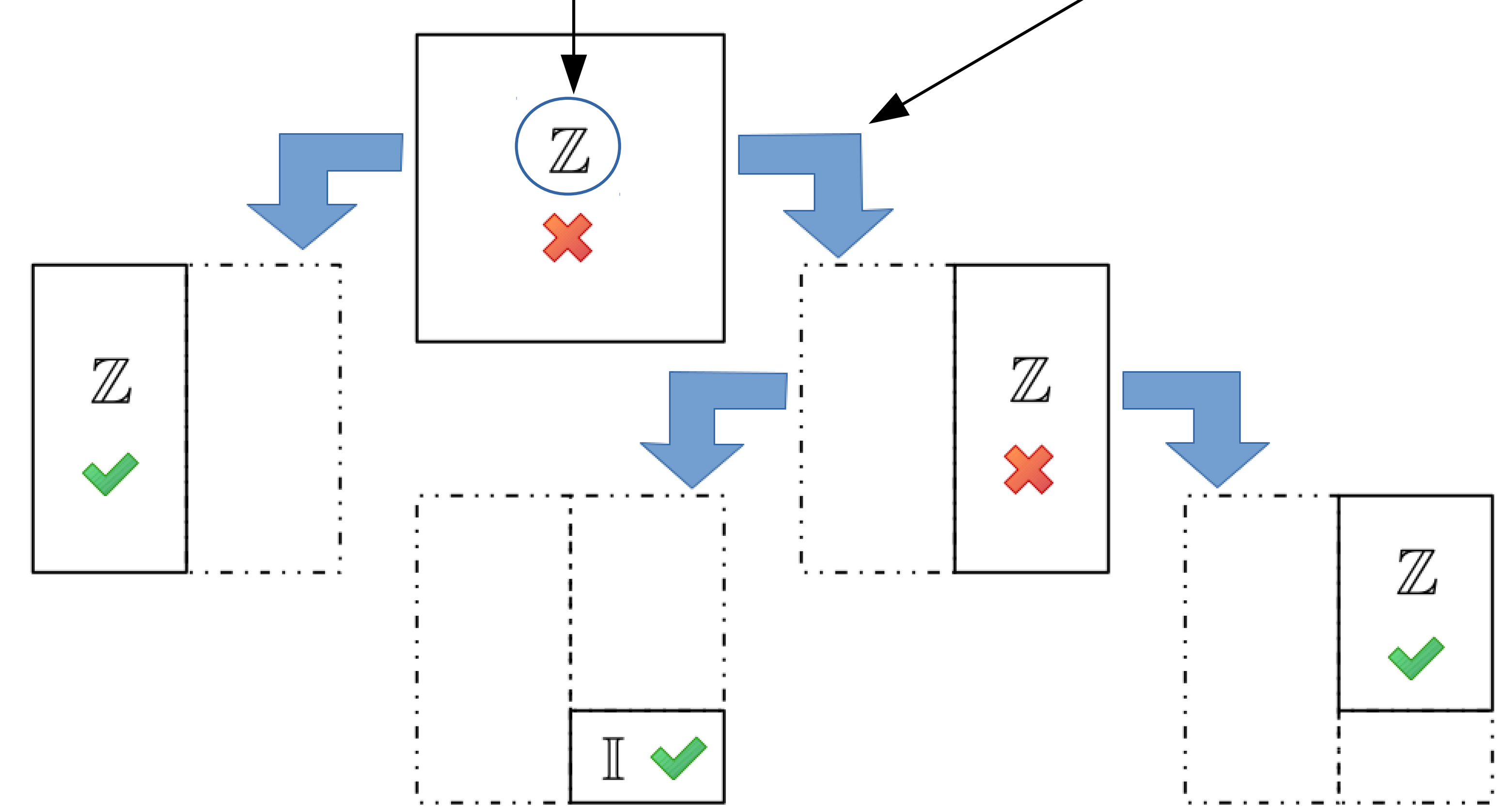


Verification policy

$$\pi_\theta = (\pi_\theta^\alpha, \pi_\theta^I)$$

How to analyze a region: chooses an abstract domain

Where to split: bisects a region into two subregions



How do we find a good verification policy?
- Neural networks are not interpretable
- There are a lot of choices to make in the policy

Policies are difficult to write by hand, so let's have the computer learn them!

We evaluate candidate policies on a set of training problems.

Evaluating a policy is expensive because we have to try to verify the robustness of several difficult training problems, so we don't have much training data.

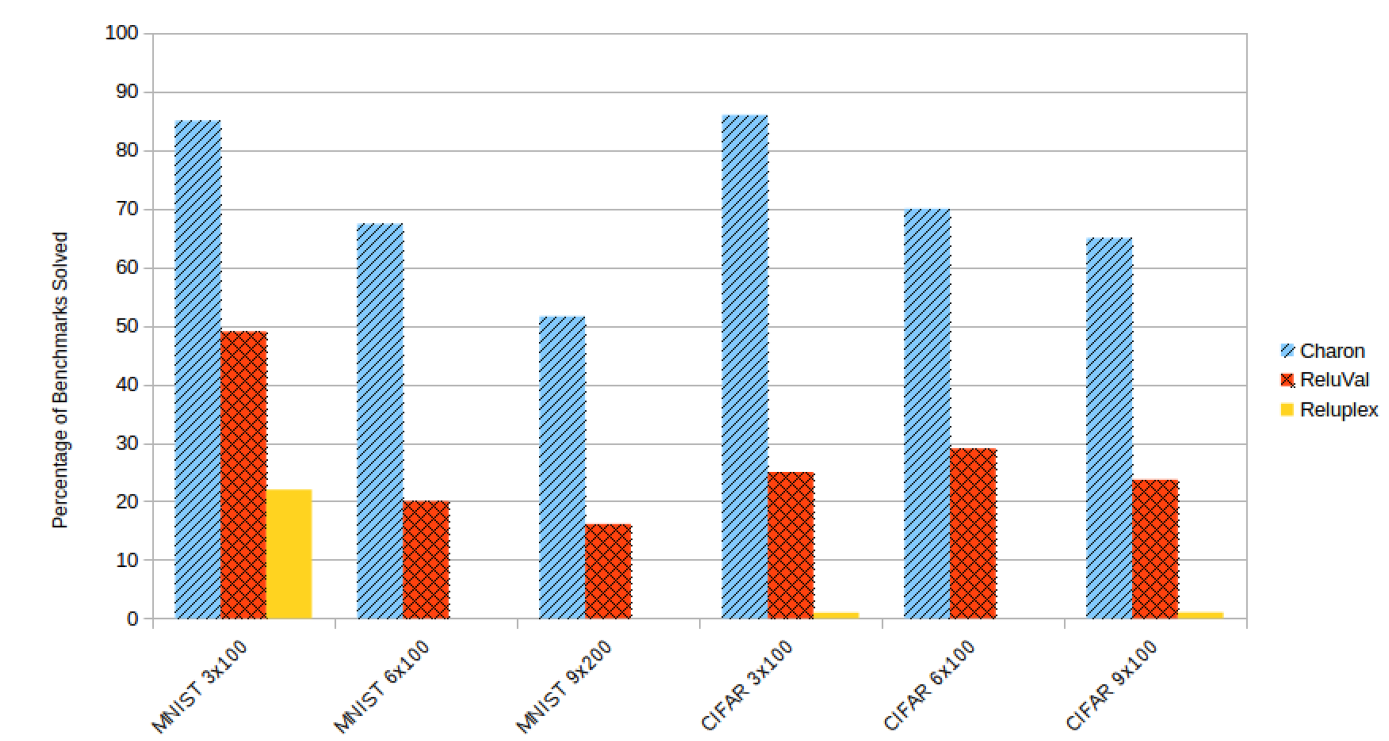
We use *Bayesian optimization,* a sample efficient machine learning framework, to learn a policy from the training problems.

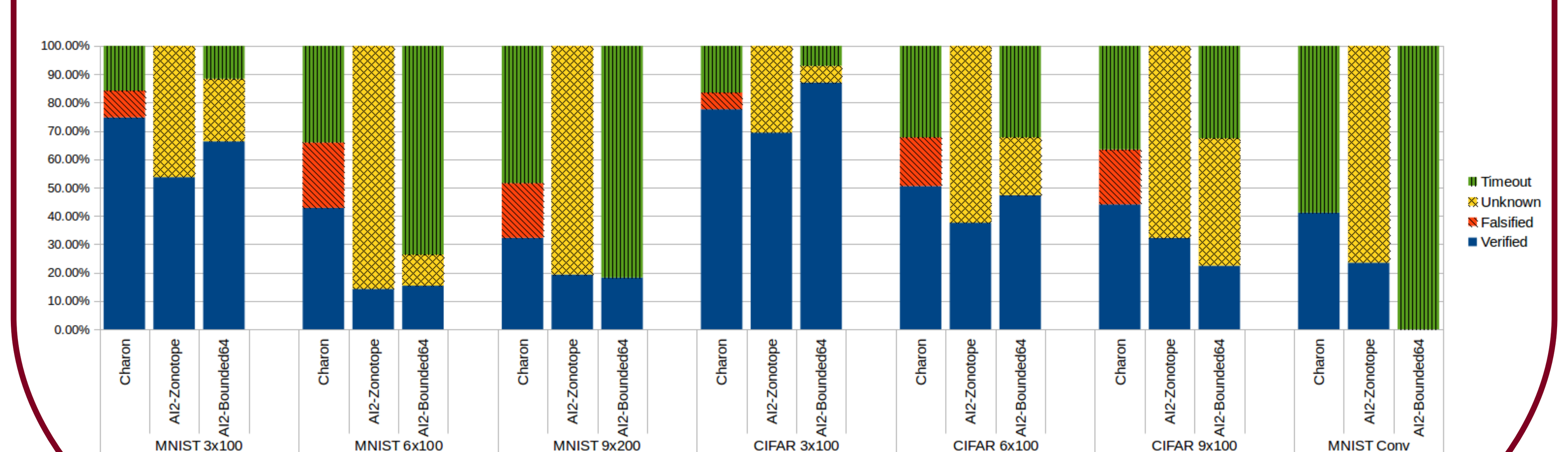
Evaluation

We implemented our approach in a tool called Charon

We evaluated Charon on 602 benchmarks across 7 networks including both convolutional and fully connected architectures

Charon substantially outperforms state-of-the-art tools for network robustness verification.





1 TEXAS The University of Texas at Austin UToPiA
2 RICE