MixTrain: Efficient Verifiably Robust Training of Neural Networks

Shiqi Wang*, Yizheng Chen*, Ahmed Abdou† and Suman Jana*
*Columbia University
†Pennsylvania State University
Formal Analysis on NN Robustness

**MNIST network**
- 98.28% test accuracy

**Test on 1000 random images**
- Symbolic Interval Analysis
- Verifiable robustness 0 at $L_\infty = 5/255 \sim 0.01$

How do we increase the verifiable robustness of neural networks?

**Verified Robust Accuracy (VRA):** percentage of test samples that are verified to be safe under a distance bound.

**Verifiably Robust Training:** 0% VRA -> over 90% VRA on MNIST, $L_\infty \leq 0.1$. 

Images Proved Safe (%) vs. $L_\infty (/255)$
Verifiably Robust Training

• The **strongest defense** against adversarial image examples.
• Provides **provable robustness** guarantee.
• Robust optimization and formal verification.
What is Verifiably Robust Training?

Regular Training
\[ \text{min}(\text{errors}) \]

Robust Training
\[ \text{min}(\text{max}(\text{errors by successful evasions})) \]

Adversarial
Underestimate Robust Loss

Verifiable
Overestimate Robust Loss
Sound Over-approximation

• Neural Network $f$
  • Input set $X$
  • The transformation $T_f$ is sound if $f(X) \subseteq T_f(X)$
  • Robustness analysis based on $T_f(X)$

• Soundness: if no successful attacks can be found by the analysis, there indeed doesn’t exist any.
New Sound Over-approximation Methods

Abstract Interpretation

Symbolic Linear Relaxation

Convex Outer Bound

SMT Solvers (Katz et al. 2017; Ehlers 2017)
Mixed-Integer Programming (Bunel et al., 2017; Tjeng et al., 2017; Cheng et al., 2017)
Semidefinite Programming (Raghunathan et al., 2018; Fazlyab et al. 2019)
Convex Relaxation (Weng et al., 2018; Mirman et al., 2018; Gehr et al., 2018; Dvijotham et al., 2018; Wong & Kolter, 2018; Wong et al., 2018)
Limitations of Verifiably Robust Training

• Expensive Sound Propagation
  • 100X slower, 1000X more memory

• Verifiable Robustness vs Accuracy Conflict

Our Work: MixTrain

Stochastic Robust Approximation

Dynamic Mixed Training
Stochastic Robust Approximation

• Verifiable Robustness
  • Training: make progress in learning
  • Testing: verify every sample

• Hyperparameter k
  • Sound over-approximation for a subset of training data
  • 1 out of 50 learns similar verifiable robustness

\[
\min( \max( \text{errors by successful evasions} ) )
\]
Dynamic Mixed Training

Objective: \( \min \left( (1 - \alpha) \text{ errors } + \alpha \max( \text{ errors by successful evasions } ) \right) \)

- Hyperparameter \( \alpha \)
  - Emphasis on Verifiable Robustness
    - Low accuracy \( \rightarrow \alpha \downarrow \)
    - High accuracy \( \rightarrow \alpha \uparrow \)
    - Training accuracy threshold
Evaluation

• Implementation with Symbolic Linear Relaxation

• 6 different networks trained over 3 benchmark datasets
  • MNIST, CIFAR, Imagenet-200

• Given a $L_p$ norm bound,
  • Estimated Robust Accuracy (ERA)
    • Under state-of-the-art attacks, e.g., Projected Gradient Descent (PGD)
  • Verified Robust Accuracy (VRA)
    • Even unknown attacks
MixTrain vs State-of-the-art Training Methods

Adversarially Robust Training [1] [2]

- Faster
  - (3x)
- Much higher VRA
  - (95.2% against 0%)
- Same ERA and ACC
  - (Within 2%)

Verifiably Robust Training [3] [4]

- Much faster
  - (15x)
- Higher VRA
  - (58% against 53%)
- Higher ERA and ACC
  - (74% against 63%)

Efficiency of MixTrain

- Wong et al.‘s method: previously the best verifiably robust training scheme.
- On CIFAR Resnet, 4.5h to reach 52% vs 9.3% VRA.
- After 12h, MixTrain has 7% higher VRA and 10% higher accuracy.
- MixTrain uses 10× less GPU memory.
First to Scale to Imagenet-200

No verifiable methods work before
- Memory problem

**MixTrain helps Wong et al.’s method work**
- MixTrain still performs better

<table>
<thead>
<tr>
<th>Method</th>
<th>Batch Time (s)</th>
<th>ACC (%)</th>
<th>VRA (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regular training</td>
<td>0.022</td>
<td>36.2</td>
<td>0</td>
</tr>
<tr>
<td>Wong et al.</td>
<td>No progress after 50 hours of training*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wong et al. + MixTrain warmup</td>
<td>3.392</td>
<td>14.4</td>
<td>5.1</td>
</tr>
<tr>
<td>MixTrain</td>
<td>0.396</td>
<td>26.2</td>
<td>19.4</td>
</tr>
</tbody>
</table>
Summary

• Two techniques for verifiably robust training.
  • Stochastic robust approximation
  • Dynamic mixed training

• MixTrain is Efficient.
  • 15× less training time and 10× less memory than state-of-the-art methods to achieve both high VRA and high accuracy.

• MixTrain is Scalable.
  • We the first to scale verifiably robust training to the ImageNet-200 dataset using off-the-shelf GPUs.