Cert-RNN: Towards Certifying the Robustness of Recurrent Neural Networks
Background

Deep Learning for Multiple Tasks

Speech Recognition

Face Recognition

Sentiment Analysis

Object Detection

Political Content Detection
Breaking Thing Is Easy

- Adversarial examples exist EVERYWHERE!

**Task:** Toxic Content Detection.  **Classifier:** LSTM.  **Original label:** 96.7% Toxic.  **Adversarial label:** 83.5% Non-toxic.

**Text:** hello how are you? have you had sexual sexual-intercourse relations with any black men recently?

**Task:** Toxic Content Detection.  **Classifier:** Perspective.  **Original label:** 92% Toxic.  **Adversarial label:** 78% Non-toxic.

**Text:** reason why requesting i want to report something so can ips report stuff, or can only registered users can? if only registered users can, then i 'll request an account and it 's just not fair that i cannot edit because of this anon block shit shiti c'mon, fucking fucking hell helled.
Defense against Adversarial Attacks

• **Empirical Defense** 😞
  - Example: Adversarial training [Madry et al. 2017]
  - Work empirically but **no theoretical guarantee**
  - Attack specific – leading to an arms race that attackers are winning

**Key Questions**
- **Guaranteed accuracy**: what is the minimum accuracy under any attack?
- **Prediction robustness**: given a prediction, can any attack change it?

• **Certified Defense**
  - **Theoretical guarantees** against all attacks within a certain threat model
  - Robustness certificate $RC(x, F, \epsilon) :$ for all $x' \in B(x, \epsilon)$ we have that $F(x) = F(x')$

ICML 2018

Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples

Anish Athalye * 1 Nicholas Carlini * 2 David Wagner 2
Robustness Certification

- The robustness property is determined by the exact decision boundary, which can be approximated by upper bound and lower bound.
- Adversarial attacks provide the asymptotic upper bound.
- The challenge is to compute the lower bound of the minimum adversarial distortions, i.e., certifying the robustness space around the input such that the model’s prediction result is consistent within the space.

Problem Statement

Given
- a neural network $N$
- a property over inputs $\varphi$
- a property over outputs $\psi$

check whether $\forall i \in I. i \models \varphi \implies N(i) \models \psi$ holds
Robustness Certification

- The basic idea to verify the robustness for a given $\ell_p$-norm perturbation space:
  - Compute the lower and upper bounds of the output units for the given perturbation space.
  - If the lower bound of the true label output is larger than the upper bounds of all other labels, the robustness for the given perturbation space is verified.

Inputs: amazing, movie

$\phi(\text{amazing}) = x^{(1)}$

$\phi(\text{great})$

$\phi(\text{outstanding})$

$\phi(\text{film})$

$\phi(\text{drama})$

$x^{(2)} = \phi(\text{movie})$

$h = A \ast \text{concat}(x^{(1)}, x^{(2)})$

Contains all possible values of $h$ subject to $x^{(1)} \epsilon \square$, $x^{(2)} \epsilon \square$

Output: $u^T \sigma(h)$

Suppose the true label is $y_2$

The robustness is verified!
Robustness Certification Methods

- **Exact Certification**
  - Satisfiability Modulo Theories (SMT) [Ehlers et al. ATVA’17, Huang et al. CAV’17, Katz et al. CAV’17]
  - Mixed-Integer Linear Programming (MILP) [Tjeng et al. ICLR ’19]
  - Accurate but usually **computationally expensive**, therefore cannot be scaled to large networks

- **Relaxed Certification**
  - Convex Polytope [Wong & Kolter ICML’18]
  - Reachability Analysis [Weng et al. ICML’18, Zhang et al. NeurIPS’18]
  - Abstract Interpretation [Mirman et al. ICML’18, Singh et al. POPL’19]
  - Efficient but **cannot provide precise robustness bounds**

However, they are almost designed for FCNs and CNNs, seldom for RNNs!
Challenges for Certifying RNNs

\[
c^{(t)} = \sigma(f^{(t)}) \odot c^{(t-1)} + \sigma(i^{(t)}) \odot \tanh(c^{(t)})
\]

\[
h^{(t)} = \sigma(o^{(t)}) \odot \tanh(c^{(t)})
\]

Cross-nonlinearity

Figure 2: The architecture of an LSTM.
Robustness Certification for RNNs

Current Works (categorized by threat model)

• Symbol/Word Substitutions \((\text{limited attackers’ ability})\)
  • Wang et al. NAACL’21
  • Dong et al. ICLR’21
  • Ye et al. ACL’20
  • Huang et al. EMNLP’19
  • Jia et al. EMNLP’19

• Embedding Perturbation \((\text{strong attackers’ ability})\)
  • POPQORN [Ko et al. ICML’19]
    • imprecise – its linear relaxations do not retain high inter-variable correlations
    • inefficient – use gradient-based optimization to compute bounding planes
    • impractical – only evaluate one single word (one input frame) perturbation

Possible? Yes!

- Precise
- Efficient
- Practical
Our Contribution

• Leveraging abstract interpretation, we propose a novel certification framework for RNNs – **Cert-RNN**, which significantly outperforms prior work in terms of both precision and efficiency.

• We conduct extensive evaluation on **four security-sensitive applications** across various network architectures to empirically validate Cert-RNN’s superiority.

• The robustness bound certified by Cert-RNN can be practically used as a meaningful quantitative metric for evaluating both the interpretability of RNNs and the provable effectiveness of various defense methods. We also demonstrate Cert-RNN’s superiority in improving the robustness of RNNs.
Method Design
Abstract Interpretation

Concrete Domain

Inputs:
- amazing movie
- great film
- amazing film
- outstanding drama
- great movie
- outstanding movie

Neuron Values:
- [0.23, 0.12, 0.48, ...]
- [0.95, 0.82, 0.11, ...]
- [0.46, 0.28, 0.73, ...]
- [0.81, 0.02, 0.01, ...]
- [0.86, 0.46, 0.23, ...]
- [0.02, 0.36, 0.57, ...]

Outputs:
- [0.45, 0.98, 0.29, ...]
- [0.85, 0.32, 0.71, ...]
- [0.26, 0.68, 0.93, ...]
- [0.11, 0.32, 0.41, ...]
- [0.76, 0.06, 0.53, ...]
- [0.22, 0.06, 0.67, ...]

Abstract Domain

Inputs:
- amazing movie
- great film
- amazing film
- outstanding drama
- great movie
- outstanding movie

Neuron Values:
- [0.45, 0.98, 0.29, ...]
- [0.85, 0.32, 0.71, ...]
- [0.26, 0.68, 0.93, ...]
- [0.11, 0.32, 0.41, ...]
- [0.76, 0.06, 0.53, ...]
- [0.22, 0.06, 0.67, ...]

Outputs:
- amazing movie
- great film
- amazing film
- outstanding drama
- great movie
- outstanding movie
Three popular numerical abstract domains

We choose **zonotope** abstract domain for the following reasons:

- **Trades off** precision and performance
- Each variable (abstract neuron) captured in an **affine** form -> **exact** for linear operations
- Allows **relating** variables through parameters

*Box* domain vs. *Zonotope* domain vs. *Polyhedra* domain
(From scalable to precise.)
Main Steps

1. A zonotope abstract domain is first defined to capture all potential adversarial inputs
2. An abstract transformer is created for each non-linear operation of the RNN
3. Propagating the zonotope through all the layers of the target RNN
4. The output zonotope of the RNN’s last layer is used to certify the robustness
Definition 4.2 Given a continuously differentiable non-linear function \( f(x_1, x_2, \ldots) \) defined in a zonotope, the zonotope approximation for \( f \) consists of two parallel planes: the lower bounding plane \( Z^L \) and the upper bounding plane \( Z^U \). We define \( Z^L \) and \( Z^U \) for any \((x_1, x_2, \ldots) \in z\) as follows:

\[
Z^L = C_1 + a_1 \cdot x_1 + a_2 \cdot x_2 + \cdots,
\]

\[
Z^U = C_2 + a_1 \cdot x_1 + a_2 \cdot x_2 + \cdots,
\]

- \( C_1, C_2, a_i \in \mathbb{R} \)
- when \( a_i = 0 (i = 1, 2, \ldots) \), the zonotope approximation returns the interval range of \( f \), i.e., \([C_1, C_2]\)

Problem Definition Given a non-linear function \( f \) and its bounding planes \( Z^L, Z^U \), its output region can be bounded by a zonotope \( z_o = a_1 \cdot z_1 + a_2 \cdot z_2 + \cdots + \frac{C_2-C_1}{2} \varepsilon_{\text{new}} \), where \( \varepsilon_{\text{new}} \) is a new error term which is introduced from the zonotope approximation for \( f \). Thus, the problem to find the tightest bound of \( z_o \) can be formalized as bellow:

\[
\min \frac{C_2 - C_1}{2}.
\]
Step 1: Input Region Abstraction

- Given an input sequence $X = [x^{(0)}, x^{(1)}, \ldots, x^{(t)}, \ldots, x^{(T)}]$, where $x^{(t)} = [x_1^{(t)}, x_2^{(t)}, \ldots, x_K^{(t)}]$ represents the $t$-th input frame.
- Based on Definition 4.1, the input frame $x^{(t)}$ is mapped to the center coefficient $\alpha_0$ of a zonotope $z$.
- For $\ell_\infty$-norm bounded attack, the adversarial perturbation of the $j$-th dimension of $x^{(t)}$ is mapped to the coefficient $\alpha_{ij}$.
Step 2: Intermediate Operation Abstraction

- **Affine Transformation Abstraction**
  - Can be exactly captured in our approximation
- **Tanh Function Abstract Transformer**
  - We propose a new abstract transformer for tanh
  - Tighter than DeepZ[Singh et al. NeurIPS’18]
Intermediate Operation Abstraction

• Sigmoid ⊙ Tanh Abstract Transformer

**Theorem 4.2.** Let \( z = \sigma(x) \cdot \tanh(y) \), where \( (x, y) \in \mathcal{Z} \subseteq [l_x, u_x] \times [l_y, u_y] \). Then, the fine-grained zonotope approximation planes in \( \mathcal{Z} \) are:

\[
Z^L = C_1 + Ax + By
\]

\[
Z^U = C_2 + Ax + By
\]

where \( A, B, C_1, C_2 \) have nine different cases as shown in Tab. 8 (deferred to Appendix B) according to the value of \( l_x, u_x, l_y \) and \( u_y \).
Intermediate Operation Abstraction

- Sigmoid ⊗ Identity Abstract Transformer

**Theorem 4.3.** Let \( z = x \cdot \sigma(y) \), where \((x, y) \in \mathcal{Z} \subseteq [l_x, u_x] \times [l_y, u_y]\). Then, the zonotope approximation planes in \( \mathcal{Z} \) are:

\[
Z^L = C_1 + Ax + By
\]

\[
Z^U = C_2 + Ax + By
\]

where \( A, B, C_1, C_2 \) have three different cases as shown in Tab. 9 (deferred to Appendix C) according to the value of \( l_x, u_x \).

<table>
<thead>
<tr>
<th>Case</th>
<th>Conditions</th>
<th>Solutions</th>
<th>Proof</th>
</tr>
</thead>
</table>
| 1    | \( l_x \geq 0 \) | \[
A = \frac{\left(\sigma(u_x) - \sigma(l_x)\right)\left(\tanh(u_y) + \tanh(l_y)\right)}{2u_x}, \quad B = \frac{\left(\sigma(u_x) + \sigma(l_x)\right)\left(\tanh(u_y) - \tanh(l_y)\right)}{2u_y}, \quad C_1 = f_{x, \sigma}(x^{**}, y^{**}) - Ax^{**} - By^{**}, \quad C_2 = f_{x, \sigma}(x^{*}, y^{*}) - Ax^{*} - By^{*}
\] | Appendix C.1 |
|      | \( u_x \leq 0 \) | In this case, we use the same method used in Case 1. |
| 3    | \( l_x < 0 \) and \( u_x > 0 \) | \[
A = \min_{\delta_x} \left\{ \frac{f_{x, \sigma}(u_x, u_y) - f_{x, \sigma}(l_x, l_y)}{u_x}, \frac{f_{x, \sigma}(u_x, u_y) - f_{x, \sigma}(l_x, u_y)}{u_x} \right\}, \quad B = 0, \quad C_1 = f_{x, \sigma}(l_x, u_y) - A l_x, \quad C_2 = f_{x, \sigma}(u_x, u_y) - A u_x
\] | Appendix C.2 |
Cert-RNN

Certifying the Robustness Bound

- Specifically, finding the largest robustness bound $\varepsilon_c$ for the input sequence with true label $c$ can be formalized as the following optimization problem:

\[
\max \quad \varepsilon_c \\
\text{s.t.} \quad \alpha_{0c} - \sum_{j=1}^{p} |\alpha_{jc} \cdot \varepsilon_j| \geq \alpha_{0i} + \sum_{j=1}^{p} |\alpha_{ji} \cdot \varepsilon_j|, \quad \forall i \neq c
\]

Algorithm 1: Computing the robustness bound.

| Result: Certified robustness bound $\varepsilon_c$ |
| Data: model $\mathcal{F}$, input sequence $X_0$, true label $c$ |
| 1 for $i$ in $T$ do |
| 2 $\varepsilon^{(t)} = 0.5$ |
| 3 for $l = 2$ to $13$ do |
| 4 $z_o = \text{CERT-RNN}(t, \mathcal{F}, X_0, \varepsilon^{(t)});$ |
| 5 if $\alpha_{c0} - \sum_{j=1}^{p} |\alpha_{cj} \cdot \varepsilon_j| \geq \alpha_{i0} + \sum_{j=1}^{p} |\alpha_{ij} \cdot \varepsilon_j|$ |
| 6 then $\varepsilon^{(t)} = \varepsilon^{(t)} + 0.5^l$; |
| 7 else $\varepsilon^{(t)} = \varepsilon^{(t)} - 0.5^l$; |
| 8 $\varepsilon_c = \min(\varepsilon^{(1)}, \varepsilon^{(2)}, \ldots, \varepsilon^{(T)})$ |
A Toy Example

Steps

1. Adversarial Input Region Abstraction

\[ x_0 = [x_1, x_2]^T = [1, -1]^T \]

2. Affine Transformation Abstraction

\[ z_{\text{input}} = [z_1, z_2]^T = [x_1 + \varepsilon \cdot x_1, x_2 + \varepsilon \cdot x_2] = [1 + \varepsilon_1, -1 + \varepsilon_2]^T \]

\[ z_h = W_h \cdot h + W_x \cdot z_{\text{input}} + b_h \]

\[ z_h = [z_{h1}, z_{h2}]^T = [-2 + 2\varepsilon_2, 3 + \varepsilon_1 - \varepsilon_2] \]

\[ z_h = \tanh(z_h) \]

\[ z_h = [z_{h1}, z_{h2}]^T = [a_1 \tilde{z}_{h1} + \frac{C_{11} - C_{12}}{2} \varepsilon_3, a_2 \tilde{z}_{h2} + \frac{C_{22} - C_{21}}{2} \varepsilon_4] = [0.2498 \tilde{z}_{h1} + 0.2685 \varepsilon_3, 0.0596 \tilde{z}_{h2} + 0.0717 \varepsilon_4]^T \]

3. Tanh Function Abstract Transformer

\[ z_0 = W_y \cdot z_h + b_y \]

\[ z_0 = [z_{01}, z_{02}]^T = [-0.4996 + 0.4996\varepsilon_2 + 0.2685\varepsilon_3, 1.1788 + 0.0596\varepsilon_1 - 0.0596\varepsilon_2 + 0.0717\varepsilon_4]^T \]

4. Affine Transformation Abstraction

\[ W_h = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, W_x = \begin{bmatrix} 0 & 2 \\ 1 & -1 \end{bmatrix}, b_h = [0]^T, h = [0]^T \]

\[ W_y = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, b_y = [1]^T \]

5. Robustness Verification for the given region

\[ \varepsilon_1, \varepsilon_2, \varepsilon_3 \in [-1, 1] \]

\[ z_{01} \in [-1.2677, 0.2685], z_{02} \in [0.9879, 1.3697] \]

The true label’s confidence value \( z_{02} \) always larger than \( z_{01} \), thus the robustness is verified for \( \varepsilon = 1 \).
Evaluation
Experimental Setting

Dataset & Models

<table>
<thead>
<tr>
<th>Dataset</th>
<th>MNIST Sequence</th>
<th>Rotten Tomatoes</th>
<th>Toxic Comment Detection</th>
<th>Malicious URL Detection</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># of Images</td>
<td>Size</td>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td>Training</td>
<td>60,000</td>
<td>28 x 28</td>
<td>23,498</td>
<td>15,564</td>
</tr>
<tr>
<td>Validation</td>
<td>/</td>
<td>/</td>
<td>3,362</td>
<td>1,562</td>
</tr>
<tr>
<td>Testing</td>
<td>10,000</td>
<td>28 x 28</td>
<td>3,016</td>
<td>1,867</td>
</tr>
</tbody>
</table>

- 8 vanilla RNNs and 9 LSTMs with different hidden units and layers for MNIST Sequence
- an RNN and an LSTM with 32 hidden units for the other three datasets

Baseline Method

- POPQORN [Ko et al., ICML’19]
Experimental Setting

Evaluation Metrics

- **Certified robustness bound**
  - The certified robustness bound of a particular sample $x$ is the maximum $\varepsilon$ for which we can certify that the model $f(x')$ will return the correct label, where $x'$ is any adversarially perturbed version of $x$ such that $\|x - x'\|_\infty \leq \varepsilon$

- **Verified accuracy**
  - The verified accuracy at $\varepsilon$ of a dataset if the fraction of data items in the dataset with certified robustness bound of at least $\varepsilon$
Effectiveness & Efficiency

Remarks

- In all cases, Cert-RNN can obtain larger robustness bounds than that of POPQORN, i.e., the result of Cert-RNN is more accurate.

- Cert-RNN is much more efficient than POPQORN in general, especially for large and complex networks.

- For Mann-Whitney U test, the p-values of all models are small enough to reject the null hypothesis, which further demonstrates the superiority of Cert-RNN.

Table 2: Evaluation results in the four scenarios, including model accuracy (Acc), mean value and standard deviation of the certified robustness bound (where a large mean implies a large robustness space), and running time.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>Acc</th>
<th>Mean</th>
<th>Std</th>
<th>Time (min)</th>
<th>Mean</th>
<th>Std</th>
<th>Time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>POPQORN</td>
<td></td>
<td></td>
<td>CERT-RNN</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>RNN-2-32</td>
<td>96.8%</td>
<td>0.0084</td>
<td>0.0037</td>
<td>0.13</td>
<td>0.0157</td>
<td>0.0077</td>
<td>6.61</td>
</tr>
<tr>
<td></td>
<td>RNN-2-64</td>
<td>94.4%</td>
<td>0.0084</td>
<td>0.0033</td>
<td>0.12</td>
<td>0.0152</td>
<td>0.0076</td>
<td>6.33</td>
</tr>
<tr>
<td></td>
<td>RNN-4-32</td>
<td>95.4%</td>
<td>0.0168</td>
<td>0.0058</td>
<td>0.30</td>
<td>0.0222</td>
<td>0.0074</td>
<td>1.72</td>
</tr>
<tr>
<td></td>
<td>RNN-4-64</td>
<td>94.8%</td>
<td>0.0034</td>
<td>0.0181</td>
<td>0.40</td>
<td>0.0056</td>
<td>0.0025</td>
<td>1.70</td>
</tr>
<tr>
<td></td>
<td>RNN-7-32</td>
<td>89.0%</td>
<td>0.0027</td>
<td>0.0016</td>
<td>0.64</td>
<td>0.0057</td>
<td>0.0025</td>
<td>4.01</td>
</tr>
<tr>
<td></td>
<td>RNN-7-64</td>
<td>92.2%</td>
<td>0.0012</td>
<td>0.0012</td>
<td>0.60</td>
<td>0.0018</td>
<td>0.0012</td>
<td>4.21</td>
</tr>
<tr>
<td></td>
<td>RNN-14-32</td>
<td>92.2%</td>
<td>0.0190</td>
<td>0.0064</td>
<td>1.44</td>
<td>0.0270</td>
<td>0.0075</td>
<td>13.44</td>
</tr>
<tr>
<td></td>
<td>RNN-14-64</td>
<td>95.8%</td>
<td>0.0089</td>
<td>0.0030</td>
<td>2.31</td>
<td>0.0166</td>
<td>0.0044</td>
<td>14.38</td>
</tr>
<tr>
<td>MNIST</td>
<td>LSTM-1-32</td>
<td>98.0%</td>
<td>0.0152</td>
<td>0.0071</td>
<td>46.78</td>
<td>0.0187</td>
<td>0.0087</td>
<td>6.26</td>
</tr>
<tr>
<td>Sequence</td>
<td>LSTM-1-64</td>
<td>99.0%</td>
<td>0.0152</td>
<td>0.0064</td>
<td>53.09</td>
<td>0.0178</td>
<td>0.0075</td>
<td>4.92</td>
</tr>
<tr>
<td></td>
<td>LSTM-1-128</td>
<td>98.0%</td>
<td>0.0143</td>
<td>0.0065</td>
<td>53.09</td>
<td>0.0184</td>
<td>0.0074</td>
<td>3.98</td>
</tr>
<tr>
<td></td>
<td>LSTM-2-32</td>
<td>96.0%</td>
<td>0.0147</td>
<td>0.0062</td>
<td>150.00</td>
<td>0.0176</td>
<td>0.0080</td>
<td>8.42</td>
</tr>
<tr>
<td></td>
<td>LSTM-2-64</td>
<td>98.0%</td>
<td>0.0145</td>
<td>0.0063</td>
<td>246.50</td>
<td>0.0167</td>
<td>0.0067</td>
<td>11.92</td>
</tr>
<tr>
<td></td>
<td>LSTM-2-128</td>
<td>97.4%</td>
<td>0.0129</td>
<td>0.0052</td>
<td>192.77</td>
<td>0.0143</td>
<td>0.0056</td>
<td>12.77</td>
</tr>
<tr>
<td></td>
<td>LSTM-4-32</td>
<td>95.0%</td>
<td>0.0093</td>
<td>0.0045</td>
<td>551.70</td>
<td>0.0095</td>
<td>0.0045</td>
<td>29.24</td>
</tr>
<tr>
<td></td>
<td>LSTM-4-64</td>
<td>97.8%</td>
<td>0.0088</td>
<td>0.0040</td>
<td>593.31</td>
<td>0.0092</td>
<td>0.0039</td>
<td>37.13</td>
</tr>
<tr>
<td></td>
<td>LSTM-7-32</td>
<td>96.6%</td>
<td>0.0054</td>
<td>0.0017</td>
<td>1522.77</td>
<td>0.0056</td>
<td>0.0015</td>
<td>90.99</td>
</tr>
<tr>
<td>RT</td>
<td>RNN</td>
<td>76.0%</td>
<td>0.0091</td>
<td>0.0049</td>
<td>1342.20</td>
<td>0.0207</td>
<td>0.0098</td>
<td>40.20</td>
</tr>
<tr>
<td></td>
<td>LSTM</td>
<td>82.0%</td>
<td>0.0091</td>
<td>0.0049</td>
<td>1342.20</td>
<td>0.0080</td>
<td>0.0026</td>
<td>2646.2</td>
</tr>
<tr>
<td>TC</td>
<td>RNN</td>
<td>90.0%</td>
<td>0.0190</td>
<td>0.0107</td>
<td>2070.60</td>
<td>0.0332</td>
<td>0.0243</td>
<td>98.40</td>
</tr>
<tr>
<td></td>
<td>LSTM</td>
<td>93.0%</td>
<td>0.0190</td>
<td>0.0107</td>
<td>2070.60</td>
<td>0.0117</td>
<td>0.0068</td>
<td>3903.60</td>
</tr>
<tr>
<td>MalURL</td>
<td>RNN</td>
<td>94.0%</td>
<td>0.0282</td>
<td>0.0132</td>
<td>2923.80</td>
<td>0.0361</td>
<td>0.0203</td>
<td>243.60</td>
</tr>
<tr>
<td></td>
<td>LSTM</td>
<td>98.0%</td>
<td>0.0282</td>
<td>0.0132</td>
<td>2923.80</td>
<td>0.0097</td>
<td>0.0044</td>
<td>9851.40</td>
</tr>
</tbody>
</table>

Table 3: Mann-Whitney U test results.

<table>
<thead>
<tr>
<th>Model</th>
<th>RNN-2-32</th>
<th>RNN-4-32</th>
<th>RNN-7-32</th>
<th>RNN-14-32</th>
</tr>
</thead>
<tbody>
<tr>
<td>p-value</td>
<td>6.93×10^{-9}</td>
<td>1.91×10^{-12}</td>
<td>2.10×10^{-27}</td>
<td>1.11×10^{-39}</td>
</tr>
</tbody>
</table>

- In all cases, Cert-RNN can obtain larger robustness bounds than that of POPQORN, i.e., the result of Cert-RNN is more accurate.

- Cert-RNN is much more efficient than POPQORN in general, especially for large and complex networks.

- For Mann-Whitney U test, the p-values of all models are small enough to reject the null hypothesis, which further demonstrates the superiority of Cert-RNN.
Effectiveness & Efficiency

Remarks

- When the number of hidden units is the same, LSTMs with less layers would be more robust.
- When the number of layers is same, LSTMs with less hidden units would be more robust.
- Too many hidden units may increase the attack surface and decrease the generalizability (i.e., have a high variance) of the model, which makes it less robust.

Figure 6: Certified robustness bound in the four scenarios. The violin plot shows the data distribution shape and its probability density, which combines the features of box and density charts. The thick black bar in the middle indicates the quartile range, the thin black line extending from it represents the 95% confidence interval, and the white point is the median.
Remarks

- The verified accuracy of Cert-RNN is much higher than that of POPQORN in most cases.
A More Threatening Scenario

Perturbing All Frames

- Cert-RNN can handle this threat model while POPQORN cannot.
- Compared with perturbing one single frame, the robustness bounds for perturbing all frames decrease to some extent.

Table 6: Results for perturbing all frames on the MNIST sequence dataset.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Std</th>
<th>Time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNN-2-32</td>
<td>0.0126</td>
<td>0.0055</td>
<td>6.8420</td>
</tr>
<tr>
<td>RNN-2-64</td>
<td>0.0130</td>
<td>0.0056</td>
<td>9.0874</td>
</tr>
<tr>
<td>RNN-4-32</td>
<td>0.0044</td>
<td>0.0044</td>
<td>0.0044</td>
</tr>
<tr>
<td>RNN-4-64</td>
<td>0.0047</td>
<td>0.0023</td>
<td>14.3441</td>
</tr>
<tr>
<td>RNN-7-32</td>
<td>0.0044</td>
<td>0.0044</td>
<td>20.5882</td>
</tr>
<tr>
<td>RNN-7-64</td>
<td>0.0017</td>
<td>0.0009</td>
<td>15.4963</td>
</tr>
<tr>
<td>RNN-14-32</td>
<td>0.0127</td>
<td>0.0036</td>
<td>31.8162</td>
</tr>
<tr>
<td>RNN-14-64</td>
<td>0.0074</td>
<td>0.0020</td>
<td>34.0596</td>
</tr>
</tbody>
</table>
Application
Certifying Adversarial Defenses

Defense Methods

- FGSM-AT (Fast Gradient Sign Method-based Adversarial Training) (Goodfellow et al. ICLR’15)
- PGD-AT (Projected Gradient Descent-based Adversarial Training) (Madry et al. ICLR’18)
- IBP-VT (Interval Bound Propagation-based Verified Training) (Gowal et al. ICCV’19)

Remarks

- Cert-RNN can provide an accurate qualitative metric to evaluate the provable effectiveness of various defenses, which would be more reliable than previous empirical metrics, e.g., the attack success rate after applying a defense method.
Improving RNN Robustness

➤ Implementation

- Our training follows [Gowal et al. CVPR’19, Mirman et al. ICML’18] – we perturb the input signal and propagate interval bounds obtained by Cert-RNN through the RNN stages.
- To train, we combine standard loss with the worst case loss obtained using interval propagation.

➤ Experimental Results

- The RNNs trained with Cert-RNN-VT achieve larger robustness bounds, outperforming the RNNs trained with IBP-VT on all three datasets. This is because the interval bounds obtained by our approximation of the tanh function is more accurate than that obtained by the IBP method.

### Table 7: Certified robustness bounds for verified robustly trained RNNs.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Original</th>
<th>IBP-VT</th>
<th>CERT-RNN-VT</th>
</tr>
</thead>
<tbody>
<tr>
<td>RT</td>
<td>0.0207</td>
<td>0.0219</td>
<td>0.0224</td>
</tr>
<tr>
<td>TC</td>
<td>0.0332</td>
<td>0.0428</td>
<td>0.0436</td>
</tr>
<tr>
<td>MalURL</td>
<td>0.0361</td>
<td>0.0702</td>
<td>0.0730</td>
</tr>
</tbody>
</table>
Identifying Sensitive Words

Remarks

• The words with smaller certified robustness bounds tend to be more important for the final prediction result, i.e., more sensitive.
Limitation & Discussion

- **Improving Zonotope Approximation**
  - Explore alternative zonotope approximations which lead to tighter robustness bounds

- **Supporting Other Norm-bounded Attacks**
  - The perturbations bounded by other norms can be considered as the subsets of $\ell_\infty$ in Cert-RNN

- **Supporting More Network Types**
  - Directly applicable to Gated Recurrent Unit (GRU) model
  - New abstract transformers for attention module in Transformers
  - The possibility for certifying sequence-to-sequence models

- **Supporting Other Threat Models**
  - Word substitution perturbation
Cert-RNN has three important advantages:

a) **Effectiveness** - it provides much tighter robustness bounds.
b) **Efficiency** – it scales to much more complex models.
c) **Practicality** - it enables a range of practical applications including evaluating the provable effectiveness for various defenses, improving the robustness of RNNs and identifying sensitive words.