Measuring Neural Net Robustness with Constraints

Osbert Bastani\textsuperscript{1}, Yani Ioannou\textsuperscript{2}, Leonidas Lampropoulos\textsuperscript{3}, Dimitrios Vytiniotis\textsuperscript{4}, Aditya V. Nori\textsuperscript{4}, Antonio Criminisi\textsuperscript{4}

\textsuperscript{1}Stanford University, \textsuperscript{2}University of Cambridge, \textsuperscript{3}University of Pennsylvania, \textsuperscript{4}Microsoft Research Cambridge

### Summary

**Motivation:** Despite having high accuracy, neural nets have been shown to be susceptible to adversarial examples, where a small perturbation to an input can cause it to become mislabeled.

**Algorithm:** We propose metrics for measuring neural net robustness and devise a novel algorithm to approximate these metrics.

**Evaluation:** We evaluate the robustness of deep neural nets with experiments on the MNIST and CIFAR-10 datasets:
- We generate more accurate estimates of robustness metrics than existing algorithms.
- We use discovered adversarial examples to fine-tune neural nets, and show that existing algorithms for improving robustness “overfit” to specific kinds of adversarial examples.

**Related literature:** Existing algorithms have been proposed for finding adversarial examples:
- Approximated as cost function minimization, and solved using L-BFGS-B (Szegedy et al. 2014).
- Fast signed-gradient heuristic (Goodfellow et al. 2015).

### Robustness Metrics

- **Problem setting:** Input space \( \mathcal{X} \subseteq \mathbb{R}^n \) and output labels \( \mathcal{L} = \{1, \ldots, L\} \).
- **Classifier:** \( f : \mathcal{X} \to \mathcal{L} \).
- **Distribution:** \( \mathcal{D} \) over inputs \( \mathcal{X} \).
- **Classifier:** \( f \) is \( (x, \epsilon) \) robust if all points \( x \) s.t. \( \|x - x\|_\infty \leq \epsilon \) have the same label as \( x \).
- **Pointwise robustness of \( f \) at \( x \):** \( \rho(f, x) = \inf \{ \epsilon \geq 0 | f \text{ is not } (x, \epsilon) \text{ robust} \} \).
- **Adversarial frequency of \( f \):** \( \Phi(f, \epsilon) = \Pr_{x \sim \mathcal{D}}[\rho(f, x) \leq \epsilon] \).
- **Adversarial severity of \( f \):** \( \mu(f, \epsilon) = \mathbb{E}_{x \sim \mathcal{D}}[\rho(f, x) | \rho(f, x) \leq \epsilon] \).

### Constraint Formulation

- **Constraint systems:**
  - Linear inequalities: \( C \equiv (w^T x + b \geq 0) \)
  - Conjunctions: \( C \equiv C_1 \lor C_2 \)
  - Disjunctions: \( C \equiv C_1 \land C_2 \)

- **Neural net \( f \) as a constraint system \( C_f(x, \ell) \):**
  - Encodes whether \( f \) outputs label \( \ell \) on input \( x \).
  - Can be constructed when \( f \) is piecewise linear (e.g., ReLUs).

- **Pointwise robustness as constrained optimization:**
  \[ \rho(f, x, \ell) = \inf \{ \epsilon \geq 0 | C_f(x, \ell) \land \|x - x\|_\infty \text{ satisfies} \} \]

### Approximation

- **Approximation:**
  - Constraint formulation is NP-hard due to disjunctions.
  - We restrict the search to a linear region around the input \( x \).
  - The resulting optimization problem is a linear program (LP).
  - The LP is very large, so we devise an abstraction-refinement constraint solving loop that significantly improves scalability.

- **Piecewise linear structure of neural nets:**

- **Generated adversarial examples:**

### Evaluation on MNIST

**Neural nets:** (i) modified LeNet, (ii) fine-tuned using baseline (Szegedy et al. 2014), (iii) fine-tuned using our algorithm.

<table>
<thead>
<tr>
<th>Neural Net</th>
<th>Acc. (%)</th>
<th>Adv. Frequency (%)</th>
<th>Adv. Severity (pixels)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>99.08</td>
<td>1.32</td>
<td>11.9</td>
</tr>
<tr>
<td>Baseline</td>
<td>99.15</td>
<td>0.99</td>
<td>6.97</td>
</tr>
<tr>
<td>Ours</td>
<td>99.23</td>
<td>1.12</td>
<td>5.03</td>
</tr>
</tbody>
</table>

### Evaluation on CIFAR-10

**Neural nets:** (i) NiN, (ii) fine-tuned using our algorithm.

<table>
<thead>
<tr>
<th>Neural Net</th>
<th>Acc. (%)</th>
<th>Adv. Frequency (%)</th>
<th>Adv. Severity (pixels)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>99.62</td>
<td>0.34</td>
<td>5.9</td>
</tr>
<tr>
<td>Baseline</td>
<td>99.65</td>
<td>0.18</td>
<td>2.9</td>
</tr>
<tr>
<td>Ours</td>
<td>99.72</td>
<td>0.12</td>
<td>2.4</td>
</tr>
</tbody>
</table>

### References
