Precise and Generalized Robustness Certification for Neural Networks

Yuanyuan Yuan, Shuai Wang, Zhendong Su

HKUST, ETH Zurich
Robustness Certification

\[ f(x) = f(x'), \forall x' \in \{\tau(x, \delta) | 0 \leq \|\delta\| \leq \|\delta_{\text{max}}\|}\]
# Robustness Certification

<table>
<thead>
<tr>
<th>$f(x) = f(x')$, $\forall x' \in {\tau(x, \delta) \mid 0 \leq |\delta| \leq |\delta_{\text{max}}|}$</th>
<th>$\forall x', f(x) = f(x')$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sound</td>
<td>✓</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$f(x) = f(x')$, $\forall x' \in {\tau(x, \delta) \mid 0 \leq |\delta| \leq |\delta_{\text{max}}|}$</th>
<th>$\exists x', f(x) \neq f(x')$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incomplete</td>
<td>✗</td>
</tr>
<tr>
<td>Complete</td>
<td>✗</td>
</tr>
</tbody>
</table>
Input Mutation

- **Pixel-level:**
  - noise
  - contrast
  - brightness
  - rotation
  - shearing

- **Geometrical:**
  - translation
  - scaling
  - foggy

- **Semantic-level:**
  - filter:
    - blur
    - artistic:
      - style 1
      - style 2
    - orientation
  - perceptual:
    - local:
    - global:
    - mouth
    - eyes
    - standing

**Diverse input mutations**
Input Mutation

Simple mutations:
- Explicit math forms
- Linear $I$

Can directly get precise input space representation

The focus of previous works

Sound & complete certification
Input Mutation

Complex mutations:
- Explicit math forms
- Non-linear /

Over-approximated input space representation

Only incomplete certification

Geometrical Filter-based
Input Mutation

Advanced mutations:
- No explicit math form
- Non-linear $I$

Never studied!

Style transfer: Perceptual-level
Overview

Prior works

Input image $x$

Mutation $\tau$

Extent $\delta \leq \|\delta_{\text{max}}\|

Input mutations due to:
1) Adversarial attacks
2) Unseen inputs

Simple mutations: Precise $I$

Certification $\phi(f)$: Complete

Complex mutation: Over-approximated $I$

Certification $\phi(f)$: Only incomplete

Advanced mutations: Infeasible

Precise $I$

Complete (low cost) & Incomplete & Quantitative

Precise:
- Deliver precise $I$

Generalized:
- Support advanced mutation
- Unified implementation
- Support conventional certification frameworks (complete/quantitative)
Motivation: Generative Model

A collection of images
Infinite images by (inter)extrapolation
Latent space

Data-driven mutations:
1) Extract mutations from diverse images
2) Represent mutations as moving directions in latent space
Motivation and Problems

\[ G(z) \]: original input
\[ G(z') \]: maximumly mutated inputs
\[ z \rightarrow z' \]: corresponds to all mutated inputs
\[ z \rightarrow z' \]: mutating direction

The problem: \( G(z) \) changes arbitrarily with \( z \)!
Two Requirements

**Continuity**: when performing mutations, \( G(z) \) changes continuously with \( z \).

**Independency**: when mutating \( G(z) \) into \( G(z') \), \( z \rightarrow z' \) should only correspond to the expected mutation.

\( z \) and \( z' \) will exclusively correspond to all mutated inputs between \( G(z) \) and \( G(z') \).
Continuity

\[ \forall z, z': \frac{1}{C}d_1(z, z') \leq d_2(G(z), G(z')) \leq Cd_1(z, z') \]

- \(d_1\): distance metric over \(z\)
- \(d_2\): distance metric over \(G(z)\)

Bound the Jacobian norm of \(G\)!
Independency

When extracting mutations, different mutations are represented as **orthogonal** directions.

When performing local mutations, projecting the mutating direction into the **non-mutating direction** of the remaining region.
Evaluation: Mutations

Findings:

The resolution of G’s training data affects the number of enabled (perceptual) mutations.

• Use higher resolution training data for the generative model.

Training data decide the enabled mutations and the maximal extent of mutations.

• E.g., To enable rotation 30°, augment the training data by rotating them 30°. But it’s unnecessary to cover all [0, 30°] to enable all rotation within [0, 30°] due to continuity.
Evaluation: Mutations

Independency  Continuity

(a) Geometrical: rotation
(b) Global-perceptual: body color
(c) Local-perceptual: opening eyes
Evaluation: Certification

Complete certification over geometrical mutations

Cost: $O((2^N)^L) \longrightarrow O((N^2)^L)$ Input to $f \circ G$ is a segment

$N$: #maximal neurons in one layer
$L$: #layers

Findings on different neural networks:

Conv vs. FC: convolution layer can enhance the robustness
Depth: deeper neural network has better robustness
Data augmentation: can also enhance the robustness
Evaluation: Certification

Quantitative certification over perceptual mutations

1) Quantifies the robustness with lower/upper bounds
2) Requires inputs are represented via segments

<table>
<thead>
<tr>
<th></th>
<th>Global</th>
<th>Local</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Orientation</td>
<td>Hair</td>
</tr>
<tr>
<td>Upper Bound</td>
<td>100%</td>
<td>98.1%</td>
</tr>
<tr>
<td>Lower Bound</td>
<td>97.6%</td>
<td>95.0%</td>
</tr>
</tbody>
</table>

More sensitive to mutating eyes

Evaluation: Certification

Quantitative certification over different mutations

1. Geometrical mutation is not a major concern;
2. Artistic-style and filter-based mutations are more effective (consistent to the texture-bias);
3. Local perceptual (may mutate key attributes) is also effective.
Summary

Input Mutation

- Diverse input mutations
- Semantic-level: original image
- Geometrical: orientation, translation, rotation, shearing
- Perceptual: artistic, global, local

Motivation: Generative Model

- A collection of images
- Infinite images by (inter)extrapolation
- Data-driven mutations:
  1. Extract mutations from diverse images
  2. Represent mutations as moving directions in latent space

Two Requirements

- Continuity: when performing mutations, \( G(z) \) changes continuously with \( z \).
- Independence: when mutating \( G(z) \) into \( G(z') \), \( z \rightarrow z' \) should only correspond to the expected mutation.

Evaluation: Certification

- Quantitative certification over different mutations:
  1. Geometrical mutation is not a major concern;
  2. Artistic-style and filter-based mutations are more effective (consistent to the texture-bias);
  3. Local perceptual (may mutate key attributes) is also effective.
Thanks!

*Contact Yuanyuan for more information.*

🌐 [https://yuanyuan-yuan.github.io](https://yuanyuan-yuan.github.io)

---

**Paper**

arxiv.org/pdf/2306.06747.pdf

**Code**

github.com/Yuanyuan-Yuan/GCert