Efficient Robustness Certificates for Discrete Data

Sparsity-Aware Randomized Smoothing for Graphs, Images and More

Aleksandar Bojchevski, Johannes Klicpera, Stephan Günnemann
Robustness Certificate

Guarantee that the prediction does not change for all $\tilde{x}$ in a ball $B_r(x)$ around the input $x$

Here $B_r(x)$ is the $L_0$ ball: the attacker can change up to $r$ bits
Given any base classifier for discrete data

Graph Neural Network
ResNet
Transformer
DNN
...

Node-level Classification
Graph-level Classification
Discretized Images
Text
Molecules (SMILES)
...

Certify a smoothed classifier w.r.t. an $L_0$ adversary
Sparsity-aware smoothing improves guarantees

Reduced complexity: $O(d^3)$ to $O(r)$

Results on Graphs, MNIST, ImageNet, ...
Certifying Graph Neural Networks

Any GNN: GCN, GAT, PPNP, GIN, ...

Perturbing both graph and node attributes

First certificate for graph-level classification
Certifying Graph Neural Networks

Different GNNs have different robustness trade-offs

- GAT
- GCN
- APPNP

Perturbing Attributes

Perturbing Graph Structure

Certified ratio

Certified radius
Randomly Smoothed Classifiers

Given:

- Any base classifier \( f : \mathcal{X} \rightarrow \mathcal{Y} \)
- Any randomization scheme \( \phi(x) \)

Certify a smoothed classifier \( g \)

\[
g(x) = \arg\max_{y \in \mathcal{Y}} \Pr(f(\phi(x)) = y) \underbrace{\text{majority vote } y^*}_{\Pr(f(\phi(x)) = y)}
\]
Certifying the Smoothed Classifiers

Majority vote $g(x)$ changes slowly

Example: $f(x) = \bullet$, but $g(x) = \circ$
Randomly Smoothed Classifiers

Goal:
Guarantee that the majority votes does not change for all $\tilde{x}$ in a ball $B_r(x)$ around the input $x$

For all $\tilde{x}$, $\Pr(f(\phi(\tilde{x})) = \bullet) > 0.5$
Choosing the Randomization Scheme $\phi(x)$

First idea: Randomly flip bits with probability $p$

$x$: $[\text{black}, \text{white}, \text{black}, \text{black}, \text{black}]$

$\phi(x)$: $[\text{black}, \text{white}, \text{black}, \text{black}, \text{white}]$

Higher $p$ leads to better guarantees

Problem: For sparse data even moderately small $p$ destroys the data
Choosing the Randomization Scheme $\phi(x)$

Sparsity aware: Treat zeros separately

$x$: 

$\phi(x)$: 

Graphs: Insert edges with $p_+$, delete edges with $p_-$

We can afford to set $p_-$ relatively high and $p_+$ relatively low without introducing too much noise in the data
Deriving the Certificate

The smoothed classifier is certifiably robust if

$$\min \Pr(f(\phi(\tilde{x})) = y^*) > 0.5$$

subject to:

$$\tilde{x} \in \mathcal{B}_r(x)$$

Find the $\tilde{x}$ that minimizes the probability of the majority vote $y^*$
Constant Likelihood Ratio Regions

The smoothed classifier is certifiably robust if

$$\min \sum_i \Pr(\phi(\tilde{x}) \in R_i) h_i > 0.5$$

subject to:

$$\tilde{x} \in B_r(x)$$

$$h_i \in [0, 1]$$

$$\sum_i \Pr(\phi(x) \in R_i) h_i = p_{y^*}$$
Constant Likelihood Ratio Regions

Observation 1: We consider w.l.o.g. only dimensions where \( x_i \neq \bar{x}_i \)

Observation 2: Number of regions is independent of \( d \)

\[
\begin{align*}
    \text{Threat model: } & B_{r_a, r_d} = \{ \bar{x} : \text{added} \leq r_a \text{ bits, deleted} \leq r_d \text{ bits} \}
\end{align*}
\]
GNNs: Setup

Threat model: Perturb either graph structure or attributes
Task: Semi-supervised node classification
Results on Node Classification

GNNs are more robust to edge deletion than edge addition

![Graph with perturbing attributes and graph structure](image-url)

- GAT
- GCN
- APPNP
Results on Node Classification

Models are more robust to edge deletion than edge addition

Average max $r_d$ radius is 6.47 with *sparse* smoothing and 1.75 without
Results on Graph Classification

First certificate for the graph-level classification task
Results on MNIST

Sparsity-aware smoothing improves the certified ratio

![Graph showing certified ratio vs radii]

- $p_+ = p_-$
- $p_+ \neq p_-$

$r$, $r_d$, $r_a$
### Other results: ImageNet

Dramatically improved runtime for the exact same (tight) certificate

<table>
<thead>
<tr>
<th>Certificate</th>
<th>Type</th>
<th>Time</th>
<th>$r = 1$</th>
<th>$r = 3$</th>
<th>$r = 5$</th>
<th>$r = 7$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cohen et al. (2019)</td>
<td>Continuous</td>
<td>&lt; 1 sec.</td>
<td>0.372</td>
<td>0.226</td>
<td>0.170</td>
<td>0.138</td>
</tr>
<tr>
<td>Dvijotham et al. (2020)</td>
<td>Discrete</td>
<td>&lt; 1 sec.</td>
<td>0.362</td>
<td>0.224</td>
<td>0.136</td>
<td>0</td>
</tr>
<tr>
<td>Lee et al. (2019)</td>
<td>Discrete</td>
<td>4 days</td>
<td>0.538</td>
<td>0.338</td>
<td>0.244</td>
<td>0.176</td>
</tr>
<tr>
<td>Ours</td>
<td>Discrete</td>
<td>&lt; 1 sec.</td>
<td>0.538</td>
<td>0.338</td>
<td>0.244</td>
<td>0.176</td>
</tr>
</tbody>
</table>
Model-agnostic, Tight, Efficient, & Sparsity-Aware Robustness Certificate

Code & Project Page: https://www.daml.in.tum.de/sparse_smoothing/
Twitter: @abojchevski