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 \textbf{not} change for all $\tilde{x}$ in a ball $\mathcal{B}_r(x)$ around the input $x$.

Here $\mathcal{B}_r(x)$ is the $L_0$ ball: the attacker can change up to $r$ bits.
Given any base classifier for discrete data

Node-level Classification
Graph-level Classification

Graph Neural Network
ResNet
Transformer
DNN
...

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

Perturbations:
- inserted edge
- deleted edge
- perturbed attribute
Certifying Graph Neural Networks

Different GNNs have different robustness trade-offs

GAT  GCN  APPNP

Perturbing Attributes

Perturbing Graph Structure
Randomly Smoothed Classifiers

Given:

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

Certify a **smoothed** classifier $g$

$$g(x) = \arg\max_{y \in \mathcal{Y}} \Pr(f(\phi(x)) = y)$$

majority vote $y^*$
Certifying the Smoothed Classifiers

Majority vote $g(x)$ changes slowly

Example: $f(x) =$ , but $g(x) =$

$Pr(f(\phi(x)) = y)$
Randomly Smoothed Classifiers

Goal:
Guarantee that the majority votes does not change for all \( \tilde{x} \) in a ball \( \mathcal{B}_r(x) \) 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 $p$

$x$: [Diagram showing flipping bits]

$\phi(x)$: [Diagram showing flipped bits]

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

$\phi(x)$: \[
\begin{array}{cccccccc}
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\end{array}
\]

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:
\[
\tilde{x} \in B_r(x)
\]

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:

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

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

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

$$\frac{\Pr(\phi(x) \in R_i)}{\Pr(\phi(\tilde{x}) \in R_i)} = c_i$$

constant
Constant Likelihood Ratio Regions

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

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

Threat model: \( B_{r_a, r_d} = \{ \tilde{x} : \text{added} \leq r_a \text{ bits, deleted} \leq r_d \text{ bits} \} \)
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

Perturbing Attributes

Perturbing Graph Structure

Certified ratio

Certified $r_a, r_d$

Certified ratio

Certified $r_a, r_d$

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 \textit{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
Other results: ImageNet

Dramatically improved runtime for the exact same (tight) certificate

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

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