Certified Defense to Image Transformations via Randomized Smoothing

Randomized Smoothing for Parametric Transformations

We generalize randomized smoothing (RS) [Cohen et al.]

\[ g(x) = \arg \max_{c \in \mathbb{D}} \mathbb{P}(f(x + \epsilon) = c) \]

for classifier \( f \), noise \( \epsilon \sim \mathcal{N}(0, \sigma^2 I) \)

Then \( g(x + \delta) = g(x) \) for \( \|\delta\|_2 \leq r_\delta \).

to randomized smoothing for parametric transformations (SPT):

\[ g(x) = \arg \max_{c \in \mathbb{D}} \mathbb{P}(f(\psi(\beta)(x)) = c) \]

for classifier \( f \), noise \( \beta \sim \mathcal{N}(0, \alpha^2 I) \)

Then \( g(\psi(x)) = g(x) \) for \( \|\gamma\|_2 \leq r_\gamma \).

requires \( \psi_\alpha \circ \psi_\beta = \psi_{\alpha+\beta} \)

Heuristic best effort defense

By applying SPT to image rotation we can obtain a heuristic defense as rotations don’t compose as required (discussed next). Here we show results for adversarial rotations of up to 30°. In the paper we investigate the tightness of the obtained robustness radius and find counterexamples.

### Certification in the Presence of Interpolation

Over a base classifier (neural network) \( f \) we construct (via RS) a classifier \( h_E \) that is robust to the rotation error. Then via SPT we construct a classifier \( g \) that is robust to transformations (e.g., rotations).

\[ \forall \beta, \gamma \in \mathbb{R} \quad \|R_{\beta+\gamma}(x) - R_\beta \circ R_\gamma(x)\|_2 \leq E \]

\[ g(\cdot): f \rightarrow h_E \rightarrow \text{stop} \]

\[ h_E(\cdot): f \rightarrow \text{stop} \pm \delta \|\delta\|_2 \leq E \]

Computing the error bound on the training set

We obtain an error bound \( E \) on the training set, that we expect to hold for samples from the data distribution \( D \) with probability \( q_E \). We use interval analysis over \( \gamma \) and sampling over \( \beta \) to derive a sound bound.

\[ q_E := \mathbb{P}_{x \sim D} \left( \max_{\gamma \in \Gamma} \|R_{\beta+\gamma}(x) - R_\beta \circ R_\gamma(x)\|_2 \leq E \right) \]

Computing individual error bounds online

For a given possible attacked \( x' = R_\gamma(x) \) we compute \( g(x') = g(x) \), without access to \( \gamma \) and \( x \). Again we use interval analysis over \( \gamma \) and sampling over \( \beta \) to obtain the bound \( E \).

\[ \max_{\gamma \in \Gamma} \|R_{\beta}(x') - R_{\beta+\gamma}(x')\|_2 \leq E \]

\[ \text{certified acc.} \quad g(x') = g(x) \text{ verified} \]

<table>
<thead>
<tr>
<th>MNIST</th>
<th>0.99</th>
<th>0.78</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIFAR-10</td>
<td>0.91</td>
<td>0.85</td>
</tr>
<tr>
<td>ImageNet</td>
<td>0.76</td>
<td>0.76</td>
</tr>
<tr>
<td>Restricted ImageNet</td>
<td>0.72</td>
<td>30.00*</td>
</tr>
</tbody>
</table>

In order to calculate the above expression we need to compute the inverse rotation \( R_\gamma^{-1}(x') \). We relax this into the following set parametrized by \( \Gamma \), over which we then chose the \( \gamma \) maximizing the outer expression:

\[ \{R_\Gamma^{-1}(x') := \{x | R_\Gamma(x) = x', \gamma \in \Gamma\} \]

We compute an overapproximating of this set by using interval analysis to invert the interpolation algorithm and obtain a lower and upper bound for each pixel in \( x \). By repeated application we can refine the result.

References


safeai.ethz.ch    github.com/eth-sri/transformation-smoothing