NNs Are Continuously Evolving!

Certified Robustness Transfer (CRT)

**Problem:** Approximating certified radius $R$ during training is slow.

**Solution:** (Knowledge Transfer) Indirectly maximize $R$ by matching outputs with a certifiably robust teacher.

**Benefit:** Process of matching outputs adds negligible overhead to the training process.

**CRT Training Objective:**

$$\mathbb{E}_{x \sim N(0, \sigma^2 I); \eta \sim \mathcal{D}} \left[ z_\phi(x + \eta) - z_\theta(x + \eta) \right]$$

We Need Certified Robustness for Critical Applications

Training Certifiably Robust NNs Is Slow

<table>
<thead>
<tr>
<th>Method</th>
<th>Training Slowdown Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>SmoothAdv (2019)</td>
<td>46.20x</td>
</tr>
<tr>
<td>MACER (2020)</td>
<td>20.86x</td>
</tr>
<tr>
<td>SmoothMix (2021)</td>
<td>4.97x</td>
</tr>
</tbody>
</table>

- Preserving certified robustness across generations of NN architectures using existing methods result in high computational costs.

We Need Certified Robustness for Critical Applications

Decision boundary of a NN that is certifiably robust at $x$ within neighborhood of radius $R$. This radius is called certified radius.

Certified Robustness Transfer (CRT)

**Problem:** Approximating certified radius $R$ during training is slow.

**Solution:** (Knowledge Transfer) Indirectly maximize $R$ by matching outputs with a certifiably robust teacher.

**Benefit:** Process of matching outputs adds negligible overhead to the training process.

**CRT Training Objective:**

$$\mathbb{E}_{x \sim N(0, \sigma^2 I); \eta \sim \mathcal{D}} \left[ z_\phi(x + \eta) - z_\theta(x + \eta) \right]$$

We Need Certified Robustness for Critical Applications

Training Certifiably Robust NNs Is Slow

<table>
<thead>
<tr>
<th>Method</th>
<th>Training Slowdown Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>SmoothAdv (2019)</td>
<td>46.20x</td>
</tr>
<tr>
<td>MACER (2020)</td>
<td>20.86x</td>
</tr>
<tr>
<td>SmoothMix (2021)</td>
<td>4.97x</td>
</tr>
</tbody>
</table>

- Preserving certified robustness across generations of NN architectures using existing methods result in high computational costs.

Table 1: When a pre-trained teacher (ResNet) is not available, we show that CRT can also be used to accelerate the process of acquiring one by using a smaller sized network as a proxy teacher. This speedup is achieved while preserving certified robustness.

```
<table>
<thead>
<tr>
<th>Method</th>
<th>Training Time (h)</th>
<th>ACR</th>
</tr>
</thead>
<tbody>
<tr>
<td>SmoothMix</td>
<td>18.98</td>
<td>0.550</td>
</tr>
<tr>
<td>CRT (ResNet20 Teacher)</td>
<td>10.07</td>
<td>0.540</td>
</tr>
</tbody>
</table>
```

**Figure 1:** Cumulative time for training newly released NNs. CRT significantly speeds up the process of training new certifiably robust NNs, whether a pre-trained teacher is available or not.

**Figure 2:** Certified robustness of new generation NNs. CRT trained NNs consistently exhibit higher robustness than their SmoothMix counterparts despite large generation gap with teacher (ResNet).