Neural Network Abstraction for Accelerating Verification

Stefanie Mühlberger¹

¹Department of Informatics, Technical University of Munich (TUM)

MOVEP 2020
Stefanie Mühlberger

- 27 years old
- Bachelor’s Degree at LMU in Mathematics
- Master’s Degree at TUM in Computer Science in Cooperation with AUDI
- Doctoral Student at TUM since March ’20
Motivation
Motivation
Motivation
Motivation
Preliminaries

Neural Network

Figure: Neural Network
Preliminaries

Neuron

Figure: One Neuron $n_i$

$$y = \phi(\sum_i w_i x_i + b)$$

$\phi$: non-linear activation function

Weights

Inputs

Bias

Outputs
Abstraction

- Create a smaller abstract model
- Keep behavioral information
- Verify properties of the abstract model
If \( x \approx y \), then \( \tilde{o} = (w_1 + w_3)x \approx o \)
Safety Guarantee

Definition (Local Robustness)
For a NN

\[ f : I \rightarrow O \]

and for each input

\[ x \in X \]

every input in an \( \epsilon \)-ball of \( x \)

\[ y \in B_\epsilon(x) \]

shall be predicted with the same value

\[ f(x) = f(y) \]
Definition (I/O-Similarity)

Neurons which compute a similar function on some set $X$ of inputs, i.e., for each input $x \in X$ to the network, they compute $\varepsilon$-close values.
Merging Several Neurons

Clustering

neuron

cluster
Approach so far...

Property $\phi$

Original Network → Activation values ↔ clustering and merging → Abstracted Network
Approach so far...

Original Network

Property $\phi$

Activation values

clustering and merging

Abstracted Network

Proof lifting
Proof Lifting

Given an abstraction $\tilde{f}$ of the original NN $f$, we can provide each (output-)neuron with an over-approximated upper bound $\tilde{u} > u$ and lower bound $\tilde{l} < l$ on its activation values.

---

Results

Network Architecture (layers × neurons/layer)

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Verification Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 × 100</td>
<td>120</td>
</tr>
<tr>
<td>7 × 100</td>
<td>200</td>
</tr>
<tr>
<td>9 × 100</td>
<td>240</td>
</tr>
<tr>
<td>6 × 300</td>
<td>443</td>
</tr>
<tr>
<td>10 × 200</td>
<td>929</td>
</tr>
<tr>
<td>6 × 500</td>
<td>819</td>
</tr>
<tr>
<td>6 × 1000</td>
<td>3,412</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Verification Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 × 100</td>
<td>197</td>
</tr>
<tr>
<td>7 × 100</td>
<td>251</td>
</tr>
<tr>
<td>9 × 100</td>
<td>366</td>
</tr>
<tr>
<td>6 × 300</td>
<td>1,441</td>
</tr>
<tr>
<td>10 × 200</td>
<td>1,527</td>
</tr>
<tr>
<td>6 × 500</td>
<td>1,211</td>
</tr>
<tr>
<td>6 × 1000</td>
<td>3,600</td>
</tr>
</tbody>
</table>

Timeout
Results
Proof Lifting

Network: $6 \times 300$
200 images to verify

<table>
<thead>
<tr>
<th>Reduction Rate (%)</th>
<th>Images Verified</th>
<th>Verification Time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>197</td>
<td>48</td>
</tr>
<tr>
<td>13.33</td>
<td>195</td>
<td>36</td>
</tr>
<tr>
<td>14.72</td>
<td>195</td>
<td>36</td>
</tr>
<tr>
<td>15.28</td>
<td>190</td>
<td>36</td>
</tr>
<tr>
<td>18.06</td>
<td>63</td>
<td>35</td>
</tr>
<tr>
<td>19.44</td>
<td>0</td>
<td>34</td>
</tr>
</tbody>
</table>
Outlook

- Apply to more general architectures
- Include other similarity measures (PCA, ...)
- Counter-example guided abstraction refinement