AI2: SAFETY AND ROBUSTNESS CERTIFICATION OF NEURAL NETWORKS WITH ABSTRACT INTERPRETATION

FEBRUARY 11, 2019
PREPARED BY: ALI HARAKEH
HOW GOOD IS YOUR NEURAL NETWORK?

HOW GOOD IS YOUR NEURAL NETWORK?

- Neural networks are not robust to input perturbations.

- **Pushing the limit**: One Pixel Attack!

- **Conclusion**: There is a need for an automated and scalable analysis to certify realistic neural networks against such input perturbations.
AUTOMATED AND SCALABLE ANALYSIS

• Used to certify large scale cyber-physical systems that use NNs.
HOW TO CERTIFY NEURAL NETWORKS?

• Given:
  o Neural Network \( N(x) \).
  o A set of inputs \( x \in \mathcal{X} \), and a property over this set \( \phi \).
  o A property over outputs \( \psi \).

• To certify a neural network, check whether:

\[
\forall x \in \mathcal{X}, \ x \in \phi \implies N(x) \in \psi.
\]

• Challenges:
  o \( \phi \) captures an unbounded set of inputs. \( \text{Over} – \text{approximation} \)
  o Traditional symbolic solutions do not scale to deep neural networks. \( \text{Abstract Interpretation} \)
• Convolutional and Fully Connected layers are just affine transforms followed by a restricted non-linearity, in this case the ReLU. Pooling layers can also be expressed this way.

• Such neural network architecture can be described with a composition of Conditional Affine Transforms (CATs).
ABSTRACT INTERPRETATION (AI) FOR AI

• Certify that neural network is robust to brightness variations:

\[ \forall x \in \mathcal{X}, x \in \phi \implies N(x) \in \psi. \]
ABSTRACT INTERPRETATION (AI) FOR AI
PROPERTIES OF ABSTRACT INTERPRETATION

• If the abstract output proves a property, we know that the property holds for all concrete values.

• Every CAT function used in classification NNs can be over-approximated by Abstract Interpretation.

• Various forms of abstract domain can be used, each resulting in a different precision at the expense of scalability.
  - Box.
  - Zonotope.
  - Polyhedron.
USE CASE OF AI2: PROVE ABSENCE OF ADVERSARIAL ATTACK

• **Step 1:** Define adversarial region around input $x$ based on the perturbation of interest. For example:

$$L_{\infty} \text{ ball} : \ Ball_\epsilon(x) = \{ y \mid \|x - y\|_{\infty} < \epsilon \}$$

• **Step 2:** Prove that there exists no image $y$ in the adversarial region where $\text{NN}(x)$ not equal $\text{NN}(y)$ using AI2.
WEAK POINTS OF AI2

- Abstract interpretation is sound but imprecise.

\[
\begin{pmatrix}
2 & -1 \\
0 & 1
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_2
\end{pmatrix}
= 
\begin{pmatrix}
y_1 \\
y_2
\end{pmatrix}
\]
WEAK POINTS OF AI2

• Abstract interpretation is sound but imprecise.

• Perturbation needs to be capturable by a set of zonotopes in a precise manner, without adding too many inputs that do not capture actual perturbations to the robustness region.

• Every new type of layer or activation function in a neural network will require the process to transform it to a CAT function, then to a concrete transformer, then to an abstract transformer.
CONCLUSION

• AI2 is an opensource software capable of certifying shallow to mediumly deep classification neural networks.

• Follow up work on AI2 provides more generalizable and faster ways to perform network certification.

• Authors’ website: http://safeai.ethz.ch/