Safe Deep Learning: Progress and Open Problems

Martin Vechev
DeepCode.ai and ETH Zurich

safeai.ethz.ch
Probabilistic + Symbolic @SRI

Symbolic Methods
- Logic
- Deduction
- Modularity
- Abstraction
- Compositionality

\[
\frac{\Gamma, \Gamma' \vdash e : \tau_1 \quad \ldots \quad \Gamma, \Gamma' \vdash e_n : \tau_n \quad \Gamma, \Gamma' \vdash e : \tau}{\Gamma \vdash \text{rec } v_1 = e_1 \text{ and } \ldots \text{ and } v_n = e_n \text{ in } e : \tau}
\]

Probabilistic Reasoning, Machine Learning
- Optimization
- Probability
- Data Driven
probabilistic + symbolic @SRI

probabilistic programming [psisolver.org]

probabilistic program

```python
def main() {
    p := Uniform(0,1);
    r := [1,1,0,1,0];
    for i in [0..r.len]
        observe (Bern(p) == r[i]);
    return p;
}
```

ML for big code [deepcode.ai]

ML-guided solvers [fastsmt.ethz.ch]

trusted artificial intelligence [safeai.ethz.ch]

grad course: https://www.sri.inf.ethz.ch/teaching/ria2018
Trusted Artificial Intelligence [safeai.ethz.ch]

grad course: https://www.sri.inf.ethz.ch/teaching/ria2018
Attacks on Deep Learning

Noisy attack: vision system thinks we now have a gibbon...

\[ x + .007 \times \text{sign}(\nabla_x J(\theta, x, y)) = x + \text{sign}(\nabla_x J(\theta, x, y)) \]

“panda”
57.7% confidence

“nematode”
8.2% confidence

“gibbon”
99.3% confidence

Tape pieces make network predict a 45mph sign

Self-driving car: in each picture one of the 3 networks makes a mistake...

Robust Physical-World Attacks on Deep Learning Visual Classification, CVPR’18

Trusted Deep Learning

Certification of Deep Learning

DL2: Deep Learning and Logic
Trusted Deep Learning

DL2: Deep Learning and Logic
DL2: Querying Neural Networks

```
find i[32, 32, 3]
where i in [0, 255],
    class (NN(i)) = 9,
    ||i - deer||_\infty < 25,
    ||i - deer||_\infty > 5
```

Find an image $i$ which gets classified to 9 (truck) where the image $i$ is within some distance of the image deer.

deer

classified as truck!
DL2: Querying Neural Networks

Find an image $i$ which gets classified to 8 with network 1 and to 9 with network 2, such that pixels in row 0:9 of image $i$ are the same as image nine.
DL2: Training Neural Networks with Logic
DL2: Bridge Logic and Differentiable Loss

**Property \( \phi \)**
- \( x - y \leq 3 \)
- \( y \leq 8 \)
- \( y \geq 2 \)
- \( x + y \leq 13 \)
- \( x + y \geq 5 \)
- \( x \geq 1 \)
- \( x - y \geq -5 \)
- \( x \leq 7 \)

**Visualized SAT**
- Formula is SAT here

**Loss** \( T(\phi) \)

**Theorem:** \( \forall x, \text{ if } T(\phi)(x) = 0 \text{ then } x \text{ satisfies } \phi \)
Attacks on Deep Learning

Noisy attack: vision system thinks we now have a gibbon...

\[ x + 0.007 \times \text{sign}(\nabla_x J(\theta, x, y)) \]

“panda” (57.7% confidence)

“nematode” (8.2% confidence)

“gibbon” (99.3% confidence)

Explaining and Harnessing Adversarial Examples, ICLR ’15

Tape pieces make network predict a 45mph sign

Robust Physical-World Attacks on Deep Learning Visual Classification, CVPR ’18

Self-driving car: in each picture one of the 3 networks makes a mistake...

Trusted Deep Learning

Certification of Deep Learning

Example: can we prove an attack does not exist? (one can plug in other safety properties)
Step 1: Define the Attacker Formally

**Space of possible attacks** will be a formal spec: a **region** around image \( x \)

Example:

L_{\infty} ball around \( x \): \( \text{Ball}_\varepsilon(x) = \{ y \mid ||x - y||_\infty < \varepsilon \} \)

**Attacker** tries to find image \( y \) in region around \( x \) where \( \text{NN}(x) \neq \text{NN}(y) \)
Step 1: Define the Attacker Formally

Attacker region:

\[ x_0 = 0 \]
\[ x_1 = 0.975 + 0.025\epsilon_1 \]
\[ x_2 = 0.125 \]
\[ \vdots \]
\[ x_{784} = 0.938 + 0.062\epsilon_{784} \]
\[ \forall i. \epsilon_i \in [0,1] \]
Step 2: Prove absence of attack

We use numerical abstract interpretation

Attacker region:

\[
\begin{align*}
    x_0 &= 0 \\
    x_1 &= 0.975 + 0.025\epsilon_1 \\
    x_2 &= 0.125 \\
    &\vdots \\
    x_{784} &= 0.938 + 0.062\epsilon_{784} \\
    \forall i. \epsilon_i &\in [0,1]
\end{align*}
\]

Captures the attack

Output constraint \( \varphi_n \):

\[
\begin{align*}
    x_0 &= 0 \\
    x_1 &= 2.60 + 0.015\epsilon_0 + 0.023\epsilon_1 + 5.181\epsilon_2 + \cdots \\
    x_2 &= 4.63 - 0.005\epsilon_0 - 0.006\epsilon_1 + 0.023\epsilon_2 + \cdots \\
    &\vdots \\
    x_9 &= 0.12 - 0.125\epsilon_0 + 0.102\epsilon_1 + 3.012\epsilon_2 + \cdots \\
    \forall i. \epsilon_i &\in [0,1]
\end{align*}
\]

All possible output distributions

Label \( i \) is possible iff: \( \varphi_n \cap \{ \forall j. x_i \geq x_j \} \neq \bot \)
## Analysis Trade-offs: Precision vs. Scalability

<table>
<thead>
<tr>
<th>Title</th>
<th>Description / Example</th>
<th>More scalable</th>
<th>Less precise</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AI²: Safety and Robustness Certification of Neural Networks with Abstract Interpretation</strong></td>
<td>Generic conceptual framework for analyzing neural networks with AI.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oakland Security &amp; Privacy, 2018</td>
<td>(with Gehr, Mirman, Drachsler-Cohen, Tsankov, Chaudhuri)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Fast and Effective Robustness Certification</strong></td>
<td>Zonotope domain with new custom abstract transformers tailored to neural networks</td>
<td>More scalable</td>
<td>Less precise</td>
</tr>
<tr>
<td>NIPS 2018</td>
<td>(with Singh, Gehr, Mirman, Pueschel)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>An Abstract Domain for Certifying Neural Networks</strong></td>
<td>New, restricted polyhedra domain with abstract transformers specifically tailored to neural networks</td>
<td>More scalable</td>
<td>Less precise</td>
</tr>
<tr>
<td>POPL 2019</td>
<td>(with Singh, Gehr, Pueschel)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Robustness Certification with Refinement</strong></td>
<td>Best of both: AI + MILP. More scalable than pure MILP solutions and more precise than pure AI (but less scalable)</td>
<td>More precise</td>
<td>Less scalable</td>
</tr>
<tr>
<td>In submission</td>
<td>(with Singh, Gehr, Pueschel)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Using AI to Train Robust Deep Learning

Idea: define abstract loss to include AI result, apply automatic differentiation on AI

<table>
<thead>
<tr>
<th>Training Method</th>
<th>Accuracy %</th>
<th>Certified %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>98.4</td>
<td>2.8</td>
</tr>
<tr>
<td>Madry et al.</td>
<td>98.8</td>
<td>11.2</td>
</tr>
<tr>
<td>DiffAI (our method)</td>
<td>99.0</td>
<td>96.4</td>
</tr>
</tbody>
</table>

Convolutional Network with 124,000 neurons, $L_\infty$ with $\varepsilon = 0.1$

Differentiable Abstract Interpretation for Provably Robust Neural Networks
ICML 2018
(with Matthew Mirman, Timon Gehr)
Challenges and Open Problems

Specification
- Typically, some norm: $L_0$, $L_1$, $L_\infty$
- How about geometric changes? Distributions?
- ∀ guarantees: unbounded number of images?

Verification
- What is a good abstraction?
- How do we leverage testing results?
- How to battle approximation loss downstream?
- Creative combinations with complete methods?

Networks
- Classification? Reinforcement Learning?
- Regression? Recurrent?
- Combinations of models?

Trade-offs
- Accuracy vs. Robustness?
- Provability vs. Accuracy?