Safe Deep Learning: Progress and Open Problems

Martin Vechev
DeepCode.ai and ETH Zurich

DEEPCODE

safeai.ethz.ch

ETH Zürich

Probabilistic + Symbolic @SRI

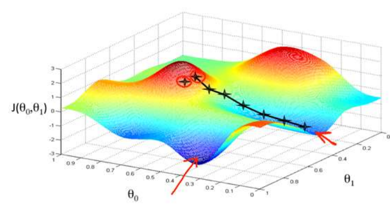
Symbolic Methods

- Logic
- Deduction
- Modularity
- Abstraction
- Compositionality

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

Probabilistic Reasoning, Machine Learning

- Optimization
- Probability
- Data Driven



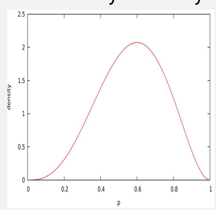
Probabilistic + Symbolic @SRI

Probabilistic Programming [psisolver.org]

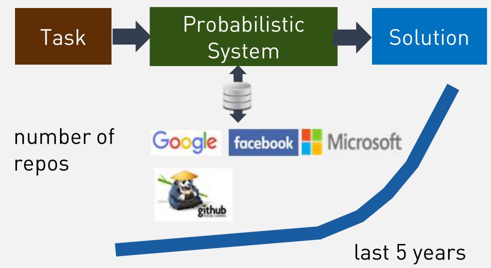
Probabilistic Program

```
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;
}
```

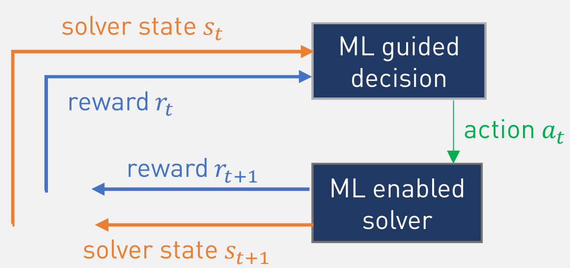
Probability Density



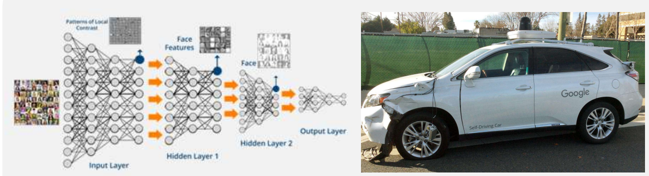
ML for Big Code [deepcode.ai]



ML-guided Solvers [fastsmt.ethz.ch/]

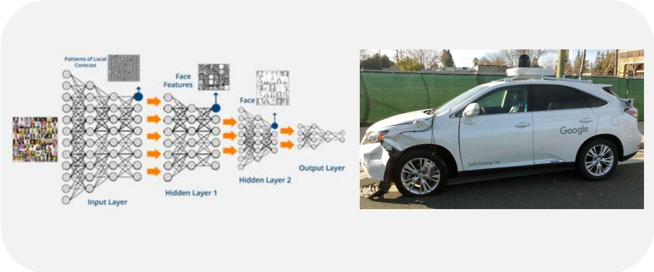


Trusted Artificial Intelligence [safeai.ethz.ch]



grad course: https://www.sri.inf.ethz.ch/teaching/riai2018

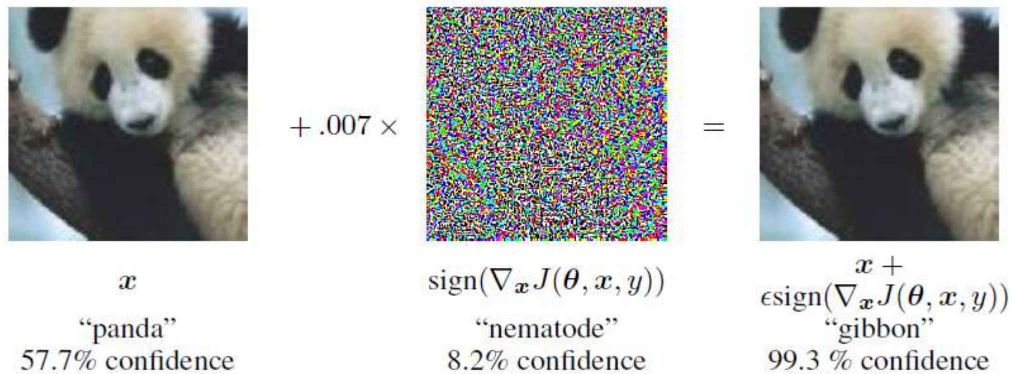
Trusted Artificial Intelligence [safeai.ethz.ch]



grad course: https://www.sri.inf.ethz.ch/teaching/riai2018

Attacks on Deep Learning

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

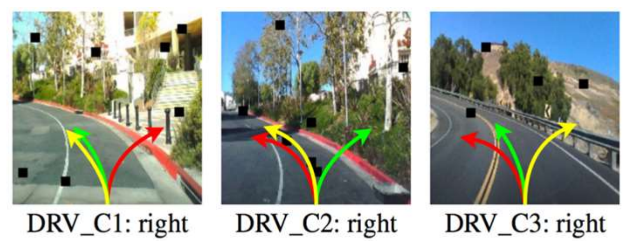


Explaining and Harnessing Adversarial Examples, ICLR '15

Tape pieces make network predict a 45mph sign

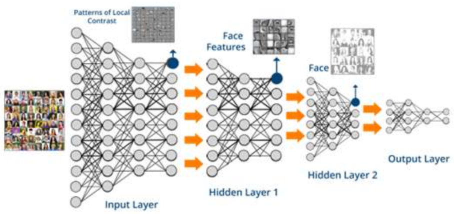
Robust Physical-World Attacks on Deep Learning Visual Classification, CVPR'18

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



DeepXplore: Automated Whitebox Testing of Deep Learning Systems, SOSP'17

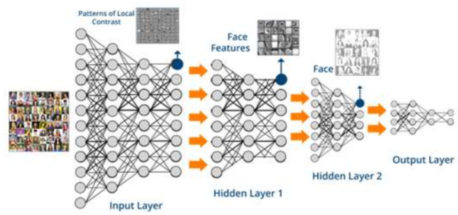
Trusted Deep Learning



Certification of Deep Learning

DL2: Deep Learning and Logic

Trusted Deep Learning



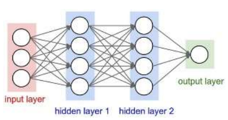
DL2: Deep Learning and Logic

DL2: Querying Neural Networks

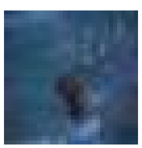
```
find i[32, 32, 3]
where i in [0, 255],
    class(NN(i)) = 9,
    ||i - deer||∞ < 25,
    ||i - deer||∞ > 5
```

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

Neural Network NN

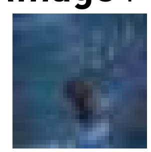


deer



Deep Learning Query Engine

image i



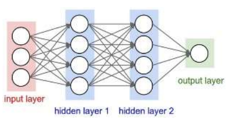
classified as **truck**!

DL2: Querying Neural Networks

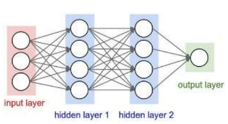
```
find i[28, 28]
where i in [0, 1],
   i[0:9,:] = nine[0:9,:],
   class(NN1(i)) = 8,
   class(NN2(i)) = 9
```

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

Network NN1



Network NN2



Deep Learning
Query Engine

image i

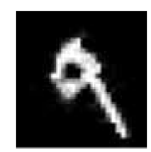
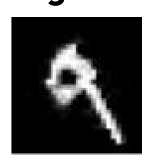
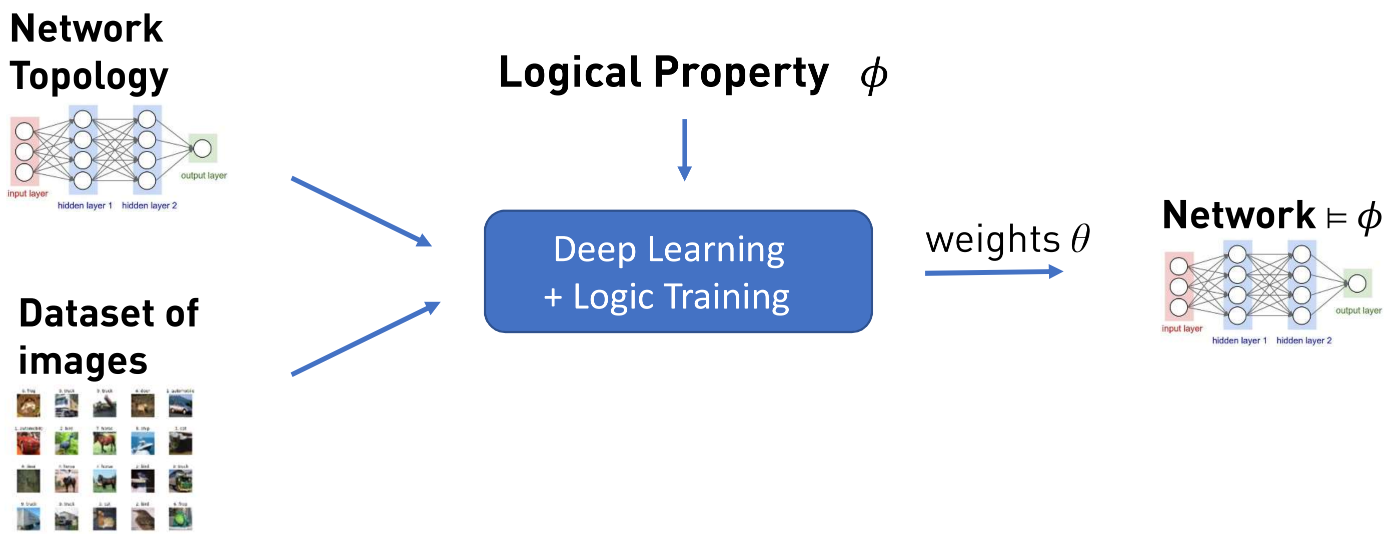


Image nine



DL2: Training Neural Networks with Logic

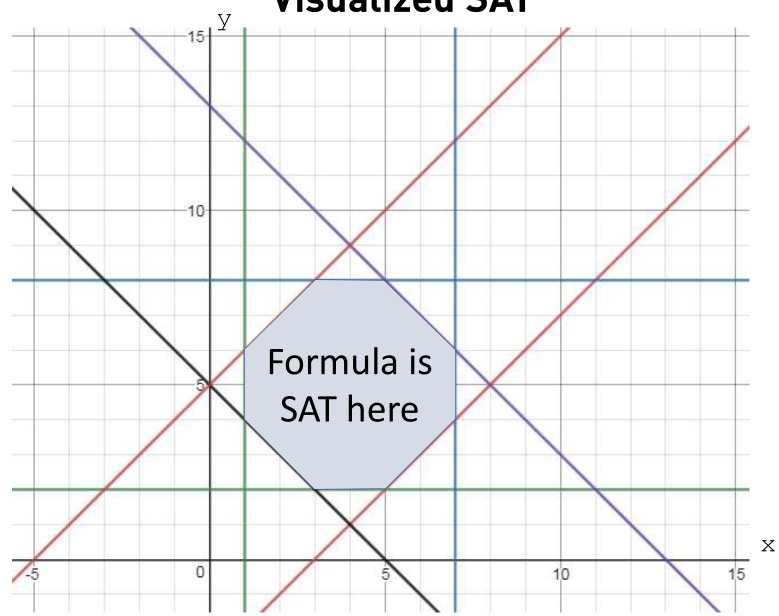


DL2: Bridge Logic and Differentiable Loss

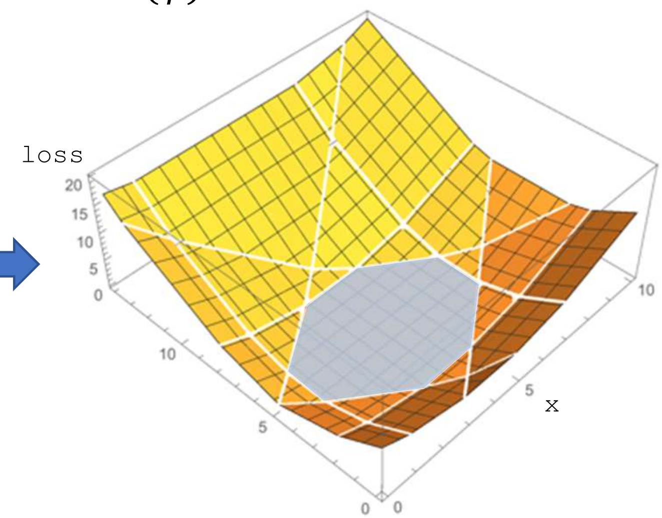
Property ϕ

```
x - y ≤ 3       ∧
y ≤ 8           ∧
y ≥ 2           ∧
x + y ≤ 13      ∧
x + y ≥ 5       ∧
x ≥ 1           ∧
x - y ≥ -5      ∧
x ≤ 7
```

Visualized SAT



Loss $T(\phi)$

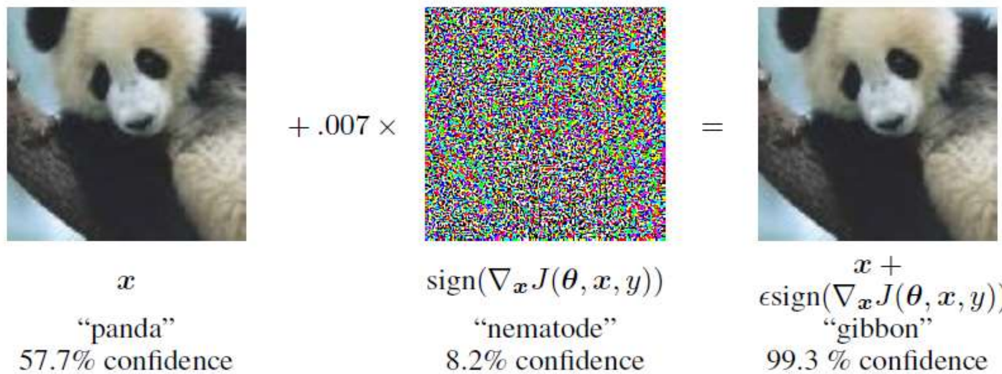


Logic to Loss Translation T

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…

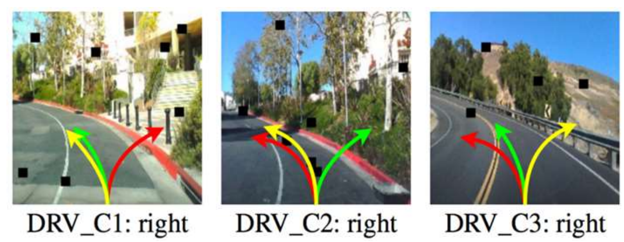


Explaining and Harnessing Adversarial Examples, ICLR '15

Tape pieces make network predict a 45mph sign

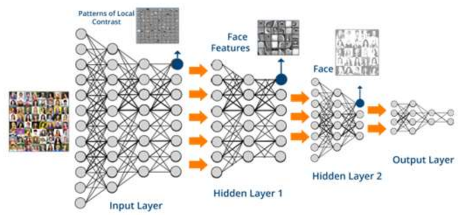
Robust Physical-World Attacks on Deep Learning Visual Classification, CVPR'18

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



DeepXplore: Automated Whitebox Testing of Deep Learning Systems, SOSP'17

Trusted Deep Learning



Certification of Deep Learning

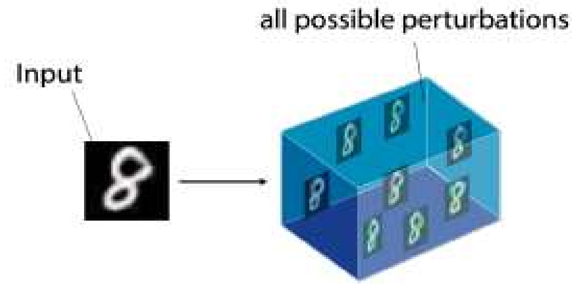
Example: can we prove an attack does not exist?
(one can plug in other safety properties)

Step 1: Define the Attacker Formally

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

Example:

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



Attacker tries to find image y in region around x where $NN(x) \neq NN(y)$

Step 1: Define the Attacker Formally

Verification

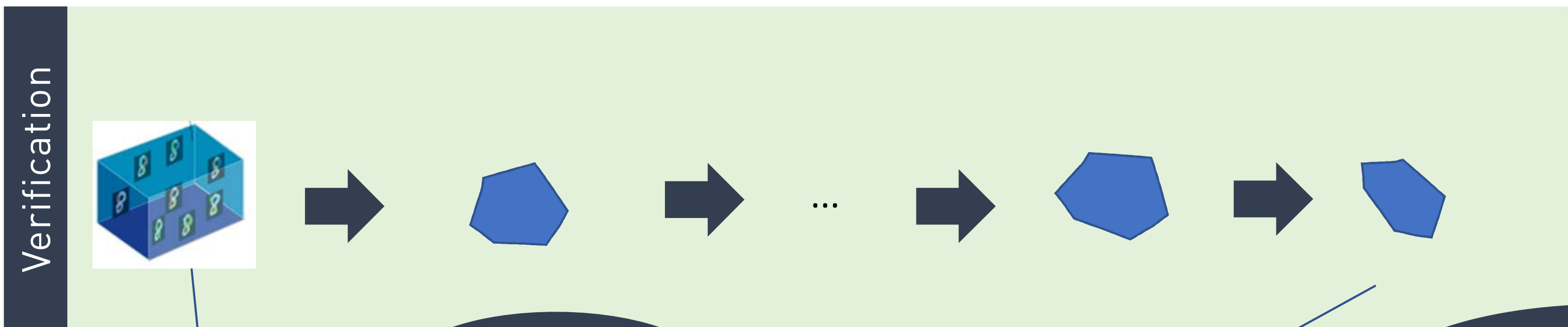
Captures the attack

Attacker region:

$x_0 = 0$
$x_1 = 0.975 + 0.025\epsilon_1$
$x_2 = 0.125$
...
$x_{784} = 0.938 + 0.062\epsilon_{784}$
$\forall i. \epsilon_i \in [0,1]$

Step 2: Prove absence of attack

We use numerical abstract interpretation



Captures the attack

Attacker region:

$x_0 = 0$
$x_1 = 0.975 + 0.025\epsilon_1$
$x_2 = 0.125$
…
$x_{784} = 0.938 + 0.062\epsilon_{784}$
$\forall i.\, \epsilon_i \in [0,1]$

All possible output distributions

Output constraint φ_n

$x_0 = 0$
$x_1 = 2.60 + 0.015\epsilon_0 + 0.023\epsilon_1 + 5.181\epsilon_2 + \cdots$
$x_2 = 4.63 - 0.005\epsilon_0 - 0.006\epsilon_1 + 0.023\epsilon_2 + \cdots$
…
$x_9 = 0.12 - 0.125\epsilon_0 + 0.102\epsilon_1 + 3.012\epsilon_2 + \cdots$
$\forall i.\, \epsilon_i \in [0,1]$

Label i is possible iff: $\varphi_n \sqcap \{\forall j.\, x_i \geq x_j\} \neq \bot$

Analysis Trade-offs: Precision vs. Scalability

AI^2: Safety and Robustness Certification of Neural Networks with Abstract Interpretation Oakland Security & Privacy, 2018 (with Gehr, Mirman, Drachsler-Cohen, Tsankov, Chaudhuri)	**Generic conceptual framework** for analyzing neural networks with AI.	
Fast and Effective Robustness Certification NIPS 2018 (with Singh, Gehr, Mirman, Pueschel)	Zonotope domain with **new custom abstract transformers** tailored to neural networks	More scalable Less precise
An Abstract Domain for Certifying Neural Networks POPL 2019 (with Singh, Gehr, Pueschel)	**New, restricted polyhedra domain** with abstract transformers specifically tailored to neural networks	
Robustness Certification with Refinement In submission (with Singh, Gehr, Pueschel)	**Best of both: AI + MILP**. More scalable than pure MILP solutions and more precise than pure AI (but less scalable)	More precise Less scalable

Using AI to Train Robust Deep Learning

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

Training Method	Accuracy %	Certified %
Baseline	98.4	2.8
Madry et al.	98.8	11.2
DiffAI (our method)	99.0	96.4

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

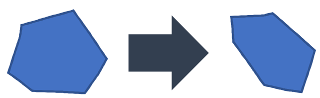
Differentiable Abstract Interpretation for Provably Robust Neural Networks
ICML 2018
(with Matthew Mirman, Timon Gehr)

Challenges and Open Problems

Specification

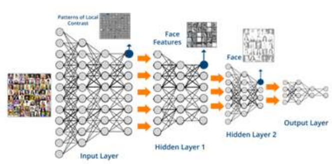
Typically, some norm: $L_0, \ L_1, \ L_\infty$
How about geometric changes? Distributions?
∀ guarantees: unbounded number of images?

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?