Building specifications for perception systems : formal proofs of deep networks trained with simulators

Julien Girard-Satabin (CEA LIST), Guillaume Charpiat (INRIA TAU),
Zakaria Chihani (CEA LIST), Marc Schoenauer (INRIA TAU)

25 février 2020

Outline

Necessity to certify deep neural networks and challenges

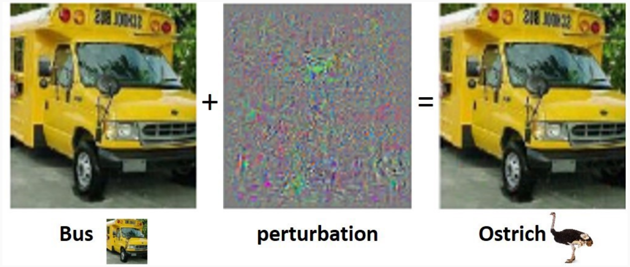
Deep neural networks are awesome. . .

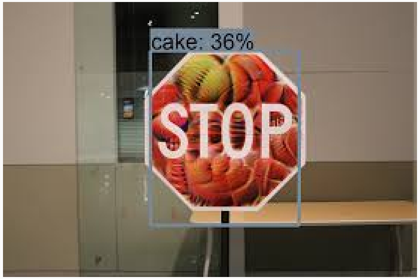
Active research community, profusion of tools, lot of industrial applications. . .

Deep neural networks are awesome. . .

Active research community, profusion of tools, lot of industrial applications. . . . . . yet they are not perfect

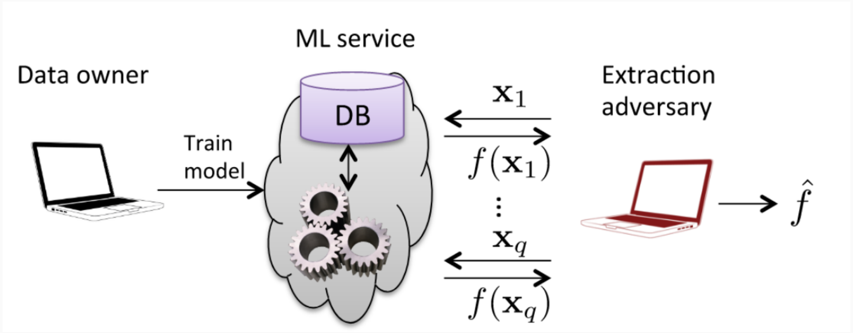
Adversarial examples (Szegedy et al. 2013)



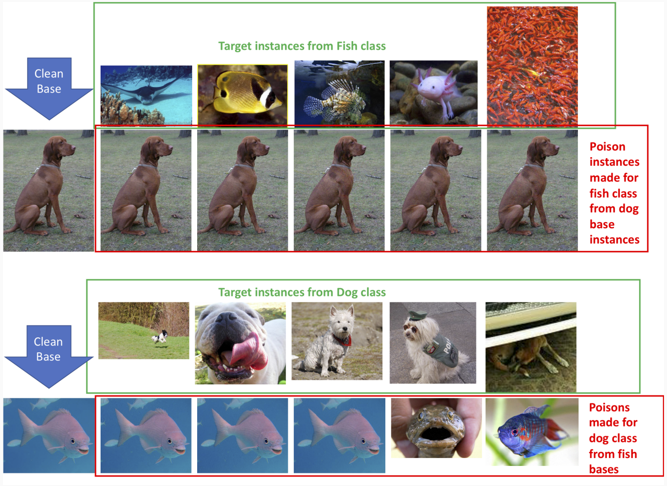


Innocuous to humans, transferable between datasets, not systematic detection method

Model theft (Tramèr et al. 2018)



Dataset poisoning (Shafahi et al. 2018)

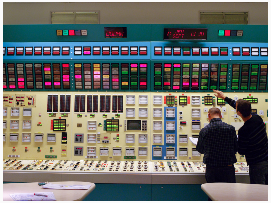

Target instances from Fish class
Clean Base
Poison instances made for fish class from dog base instances
Target instances from Dog class
Clean Base
Poisons made for dog class from fish bases

Context : critical systems

A critical system is a system whose failure may cause physical harm, economical losses or damage the environment

Context : critical systems

A critical system is a system whose failure may cause physical harm, economical losses or damage the environment



How to bring confidence in software systems ?

Goal : guarantee that the system respects a *safety specification*

ϕ : *an autonomous car will not run over pedestrians*

What about tests ?

Test on real environment	real conditions	cumbersome, potentially hazardous, non exhaustive
Test on virtual environment	can be automated, easy to integrate in existing workflow	non exhaustive, biased towards success

And more (fuzzing. . .)

Useful technique, widely used, enough for most of use cases

Sometimes, tests alone are not enough !

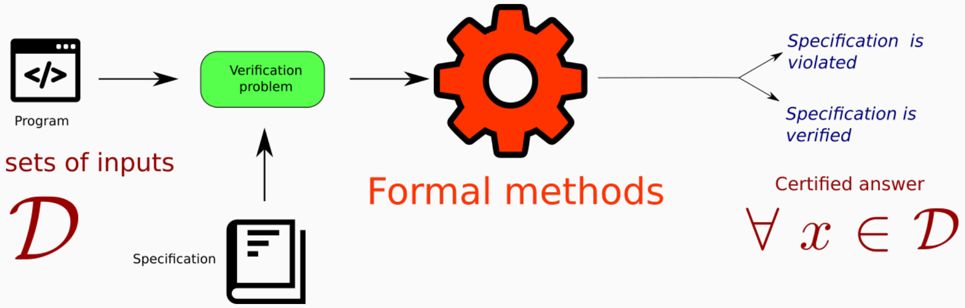
Claim	Discussion
"A car drove 5,472km, 99% in autonomous mode"[1]	If it translate to a failure rate, 10^{-2} , insufficient compared to requirements in other critical systems (about 10^{-6} in aerospace)
"Our test cases are exhaustive"	Testing sets tend to be biased towards "normal" operation (accidents are rare)[2]

1. https ://www.wired.com/2015/04/delphi-autonomous-car-cross-country/
2. https ://arstechnica.com/cars/2019/05/feds-autopilot-was-active-during-deadly-march-tesla-crash/

Introducing formal methods

- Studied in the academics since 1930 ($\lambda-$calculus, Church, Turing)
- Different techniques : abstract interpretation (Cousot and Cousot 1977), SAT/SMT (Davis and Putman 1960 ; Tinelli 2009), deductive verification (Coquand 1989), etc.
- Used in industrial settings such as aerospace, automated transports, energy to *formally* certify

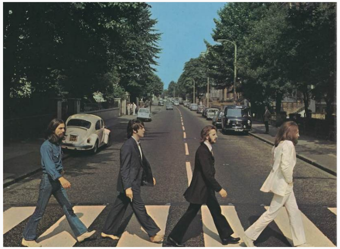
Key points

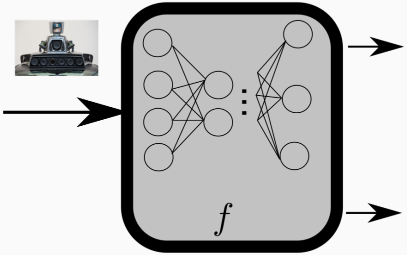


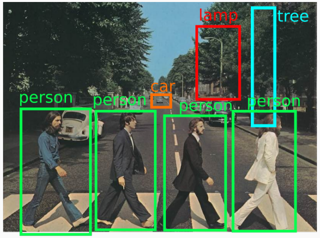
Work on domains $\mathcal{D}$ of inputs (global properties)

Answer is sound, formally guaranteed by mathematical logic

Case study : a self-driving car perception unit

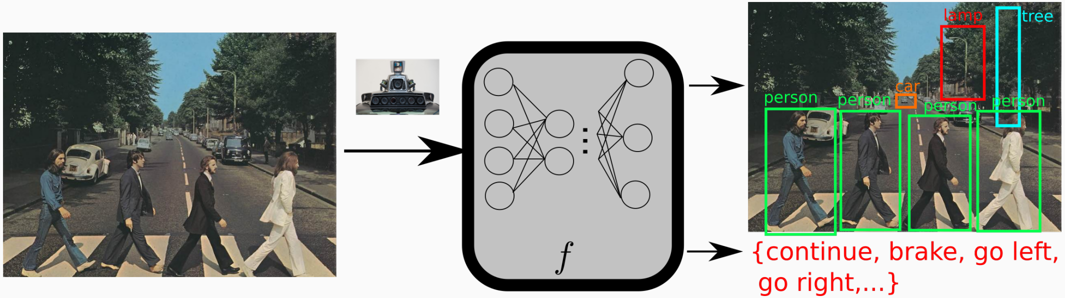






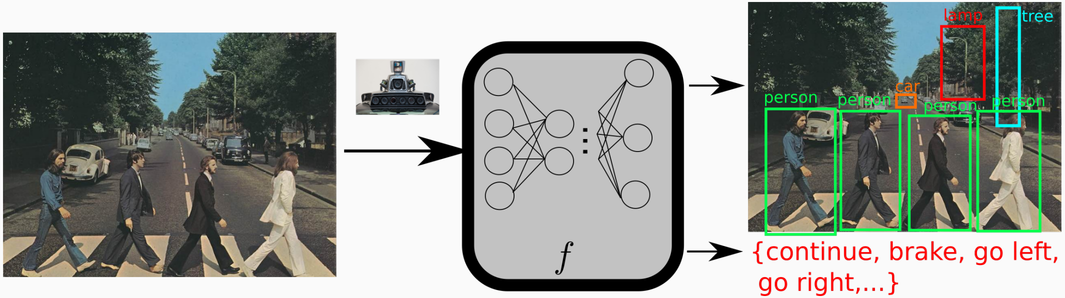
{continue, brake, go left, go right,...}

Case study : a self-driving car perception unit



Dream property ϕ : *the autonomous car never run over pedestrians*

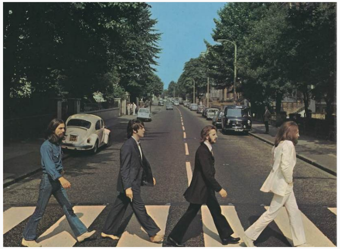
Case study : a self-driving car perception unit

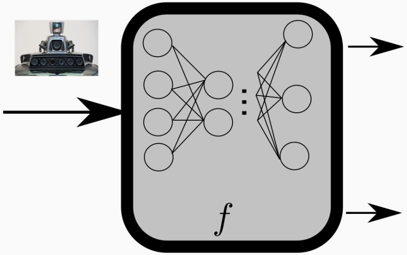


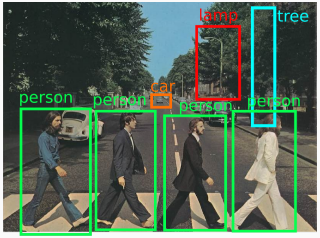
Dream property ϕ : *the autonomous car never run over pedestrians*

no *formal characterization* of what a pedestrian is !

Case study : a self-driving car perception unit







{continue, brake, go left, go right,...}

Dream property ϕ : *the autonomous car never run over pedestrians*

no *formal characterization* of what a pedestrian is !

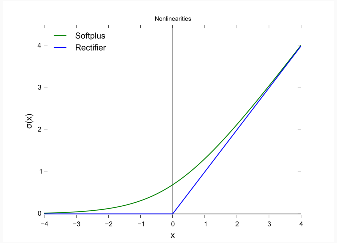
Lack of formal definition on inputs prevents from formulating interesting safety properties

It’s hard to use formal methods on deep learning

Classical software	Machine learning
Explicit control flow	Generated control flow
Explicit specifications	Data-driven specifications (lack of generality)
Abstractions and well known concepts	Very few abstractions and reusability
Documented and understood vulnerabilites	Flaws without systematic characterization

Some differences between classical software and machine learning

Another difficulty : performance of verification tools



2 cases per ReLU node for the solvers

Several million ReLU nodes $\rightarrow$ $2^{O(10^6)}$ case splits

Combinatory explosion (if done naively)

Deep learning verification : a review

Local properties : adversarial robustness

For a *given* input **x**, a classification function f, an adversarial perturbation δ :

find delta

satisfying

classifier misclassification

such that

perturbation stays below a certain threshold

Local properties : adversarial robustness

For a *given* input **x**, a classification function f, an adversarial perturbation δ :

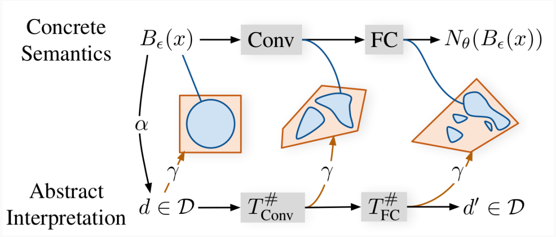
find delta

satisfying $f(x) \neq f(x + \delta)$

such that $\|\delta\|_p \leq \varepsilon$

DiffAI/DeepPoly (Gehr et al. 2018, Singh et al. 2019)

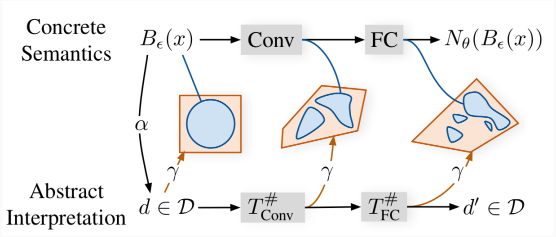
1. *abstract* the network
2. *propagate* perturbations
3. *assess* robustness properties
4. *learn to minimize* adversarial loss



Improve adversarial robustness on 100 samples from CIFAR-10 from 0 to 80%, $\varepsilon = 8/255$, 3 hidden layers, convolutional network

DiffAI/DeepPoly (Gehr et al. 2018, Singh et al. 2019)

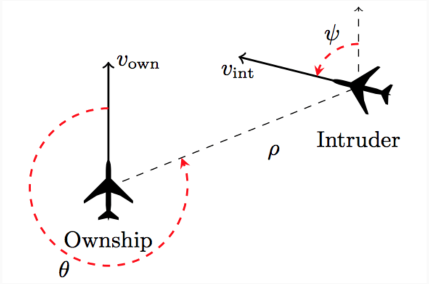
1. *abstract* the network
2. *propagate* perturbations
3. *assess* robustness properties
4. *learn to minimize* adversarial loss



Improve adversarial robustness on 100 samples from CIFAR-10 from 0 to 80%, $\varepsilon = 8/255$, 3 hidden layers, convolutional network

Scalable verification, but local properties

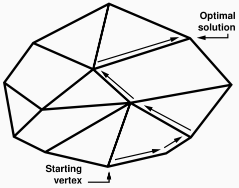
Global properties : ACAS-Xu



If the intruder is distant and is significantly slower than the ownship, the score of a COC advisory will always be below a certain fixed threshold. Bounds : $\rho \geq 55947.691$, $v_{own} \geq 1145$, $v_{int} \leq 60$

Critical system

ReLuPlex and Marabou (Katz et al. 2017, Katz et al. 2019)



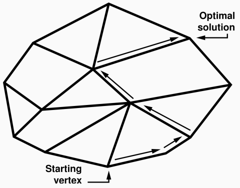
Core of most SMT solvers working with number values

Modified to lazily evaluate ReLUs

Exact verification of several properties on a ACAS-Xu implementation

Global properties

ReLuPlex and Marabou (Katz et al. 2017, Katz et al. 2019)



Core of most SMT solvers working with number values

Modified to lazily evaluate ReLUs

Exact verification of several properties on a ACAS-Xu implementation

Global properties

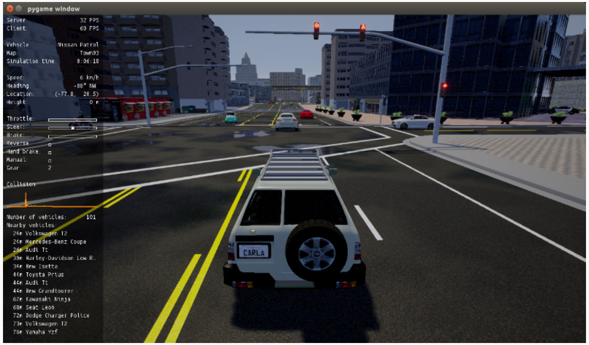
Assumes perfect plane detection beforehand

How do we verify perception ? What is an intruder ?

Verification of perception models trained with simulators

Example of simulator

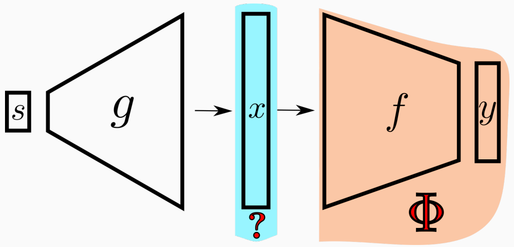
Industry relies more and more on simulators to generate scenarios to train and evaluate deep learning models



```
Server:                 32 FPS
Client:                 60 FPS

Vehicle:          Nissan Patrol
Map:                     Town03
Simulation time:        0:06:10

Speed:                   6 km/h
Heading:                -88° NW
Location:        (-77.0,  28.5)
Height:                     0 m

Throttle:
Steer:
Brake:
Reverse:       ☐
Hand brake:    ☐
Manual:        ☐
Gear:          2

Collision:

Number of vehicles:        101
Nearby vehicles:
  24m Volkswagen T2
  24m Mercedes-Benz Coupe
  26m Audi Tt
  30m Harley-Davidson Low R..
  34m Bmw Isetta
  44m Toyota Prius
  44m Audi Tt
  44m Bmw Grandtourer
  67m Kawasaki Ninja
  68m Seat Leon
  72m Dodge Charger Police
  73m Volkswagen T2
  76m Yamaha Yzf
```

Screenshot from the CARLA open source simulator

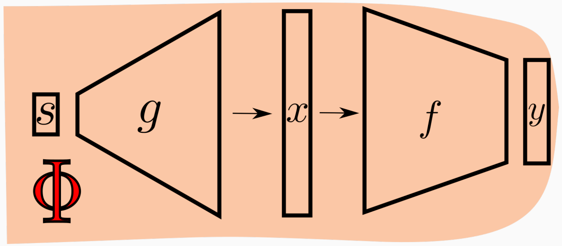
Simulators as data providers



- s : parameters (obstacles, weather conditions...)
- g : simulator
- x : perceptual input (images)
- f : model
- y : decision output (brake...)
- ϕ : "$\forall\ x$ that contains a pedestrian, do not roll over it"

How to formulate ϕ ? What is an image x with a pedestrian ?

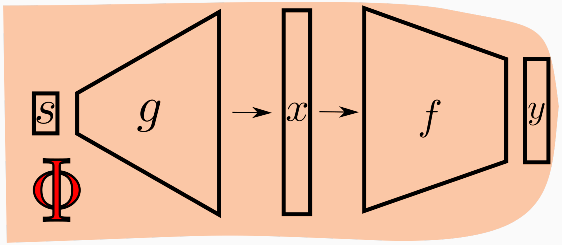
Reformulation of our verification problem



Modify the verification problem formulation to include g and s

ϕ now encompasses s and can now be expressed : *For all values of* s *that are translated by* g *as the presence of pedestrians into* x*, do not run over those pedestrians*

Reformulation of our verification problem

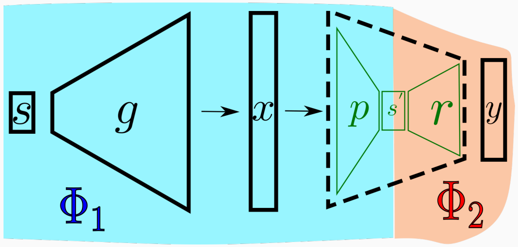


Modify the verification problem formulation to include g and s

ϕ now encompasses s and can now be expressed : *For all values of s that are translated by g as the presence of pedestrians into x, do not run over those pedestrians*

We now have a property to verify a perceptive unit !

Refinement : splitting perception and reasoning

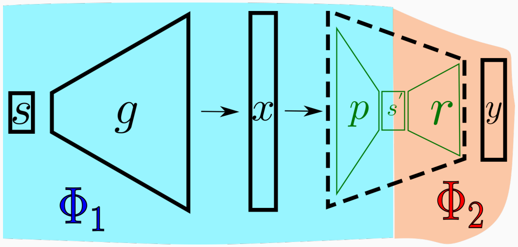


f splits in *perception* and *reasoning*, p learns s

ϕ_1 on p : guarantee of no information loss : *reconstruct s from x*
$s' = s \;\forall\; s \rightarrow p \circ g = Id$

ϕ_2 on r : do not kill pedestrians (assuming perfect perception)

Refinement : splitting perception and reasoning



f splits in *perception* and *reasoning*, p learns s

ϕ_1 on p : guarantee of controlled information loss : *reconstruct s from x*
$s^{'} \simeq s \; \forall \; s \rightarrow p \circ ||g - id < \varepsilon||$

ϕ_2 on r : do not kill pedestrians (assuming perfect perception)

How to achieve that concretely ?

How to express ϕ, g, f, $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{S}$?

How to achieve that concretely ?

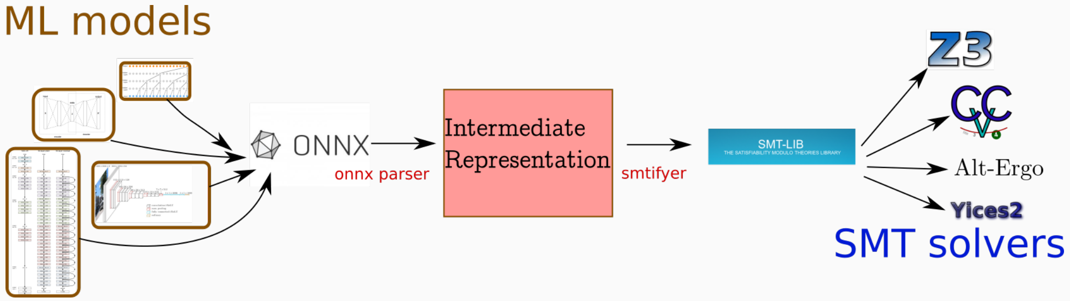
How to express ϕ, g, f, $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{S}$?

ONNX

SMT-LIB
THE SATISFIABILITY MODULO THEORIES LIBRARY

SMTLIB : Tinelli et al., 2017, https ://onnx.ai/

Toolkit to translate deep neural networks into SMTLIB



High-level workflow

From all mainstreams deep learning frameworks to all mainstreams SMT solvers

Proof of concept and future works

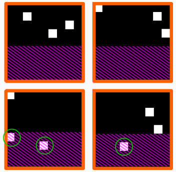
Synthetic experiment : a simple self driving car perceptive unit

Train a simple model to output a single command directive if a simplified input is in a pre-defined danger zone

Synthetic experiment : a simple self driving car perceptive unit

Train a simple model to output a single command directive if a simplified input is in a pre-defined danger zone

$s =$ (position of obstacles)



x

output scalar (obstacle detected if > 0)

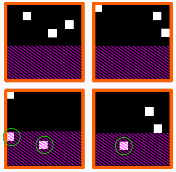
y

Network has 16 neurons, 2 hidden layers

Synthetic experiment : a simple self driving car perceptive unit

Train a simple model to output a single command directive if a simplified input is in a pre-defined danger zone

$s =$ (position of obstacles)



x

output scalar (obstacle detected if > 0)

y

Network has 16 neurons, 2 hidden layers

We prove the given trained network will *never fail*

Translation of model

```
        . . .
(declare-fun l_1.weight38 () Real)
(assert (= l_1.weight38 (/ (- 13947381) 1208925819614629174706176)))
(declare-fun l_1.weight37 () Real)
(assert (= l_1.weight37 (/ (- 405697 ) 2251799813685248)))
(declare-fun l_1.weight36 () Real)
        . . .
```

Part of the network's translation

Future work

1. Include noise and incomplete reconstruction in the framework
2. Add rewriting rules
3. Release and enhance the toolkit
4. Add a systematic representation of the simulator
5. Integration of state-of-the art verification tools

Questions ?

Any questions ?

contact for the paper : julien.girard2@cea.fr