Safety verification of AI models and other AI works at ESTAS lab

Pierre-Jean Meyer

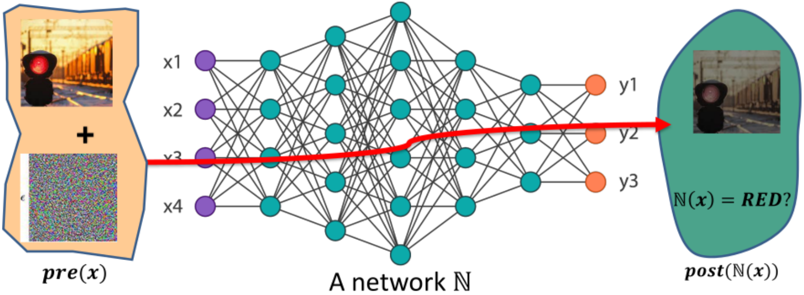
Université Gustave Eiffel

18^{th} of September 2024

Neural network verification

Context: image classification for railway systems

- Neural networks often sensitive to noise
- Need for formal verification of its safety



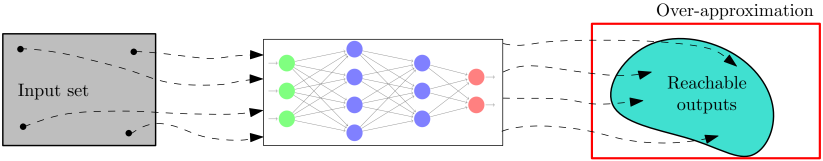
Safety specification: input-output property

- Input set: original image + set of allowed disturbances
- Output set: classifications identical to the original image

Structure of NN verifiers

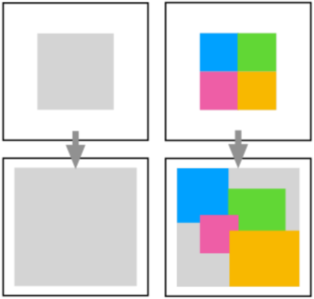
Reachability analysis

- What is reachable from the input set ?
- **Over**-approximation of the true reachable set



Iterative refinement

- Split input set for more accuracy
- Reachability analysis for each subset



Outline

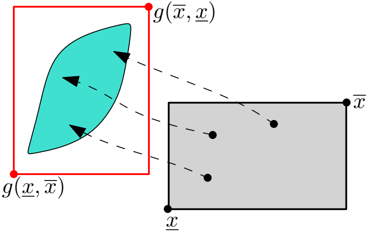
Mixed-monotonicity reachability analysis

$$y = f(x), \; x \in X, \; y \in Y$$

Definition (Mixed monotonicity)

Function f is mixed-monotone is there exists $g : X \times X \to Y$ such that

- $g(x, \hat{x})$ is increasing with x
- $g(x, \hat{x})$ is decreasing with $\hat{x}$
- $g(x, x) = f(x)$



- Only 2 evaluations of g
- Needs only bounds on the derivative f'
- Applicable to any continuous activation function
  - $\to$ any neural network

Reachability comparisons

	Mixed-mono	Symbolic interval propagation
Generality	Continuous AF	ReLU, Sigmoid
Complexity	Polynomial	Linear
Tightness	Complementarity	

Future direction: combine with iterative refinement algorithm

Reachability comparisons

	Mixed-mono	Symbolic interval propagation
Generality	Continuous AF	ReLU, Sigmoid
Complexity	Polynomial	Linear
Tightness	Complementarity	

Future direction: combine with iterative refinement algorithm

Neural networks with uncertain parameters (weights, biases)
$\rightarrow$ straightforward extension for mixed-monotonicity

	Mixed-mono	Symbolic interval propagation
Generality	Continuous AF	ReLU, Sigmoid
Complexity	Polynomial	Exponential
Tightness	Tighter	Looser or does not run

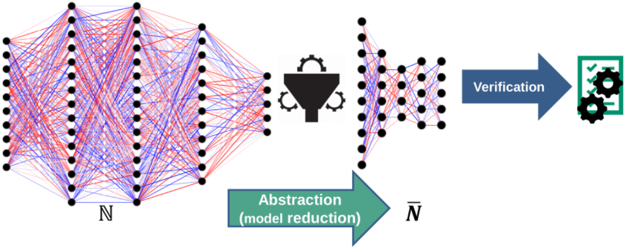
Future direction: combine with NN model reduction (next slides)

Outline

NN model reduction

Scalability of NN verification algorithms

- Limitation : high complexity due to the large size of neural networks
- Solution : model reduction before the verification step
  $\rightarrow$ need to **over-approximate** the behaviors of the original NN



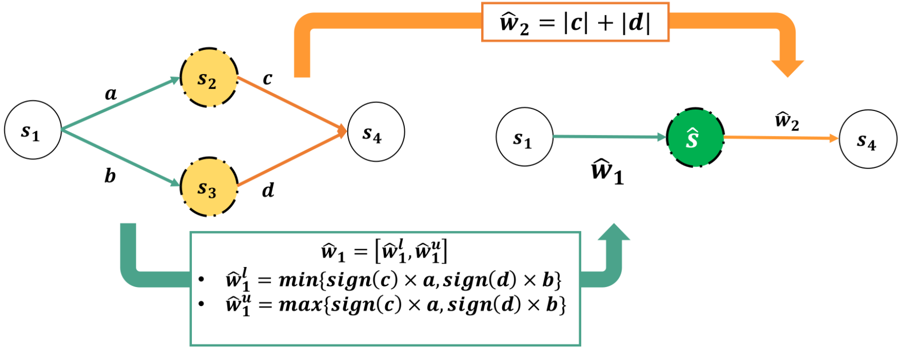
Trade-off complexity/precision

- Reduced verification computation time
- At the cost of more conservative results

Merging two nodes

For **odd and increasing** activation functions (TanH)

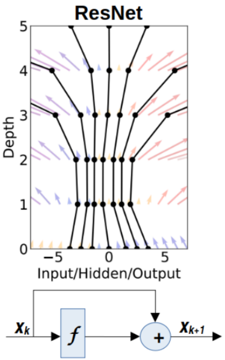
- Outgoing edges: sum of absolute values
- Incoming edges: interval bounds



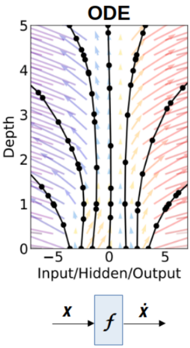
Future direction: combine this interval NN with NN reachability

Outline

Neural Ordinary Differential Equation



$$x_{k+1} = x_k + f(x_k)$$



$$\frac{d}{dt} x = f(x)$$

Infinitely more intermediate layers

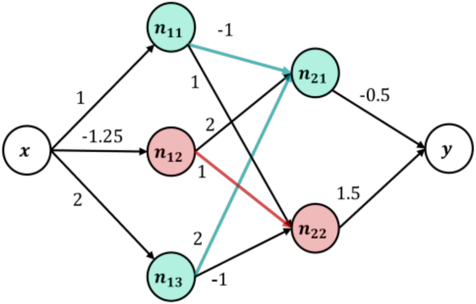
- unchanged final "depth"
- unchanged final result

Research directions

- Verification of nODE using reachability analysis of continuous systems
- Formal relationships between neural ODE and NN models

Outline

Monitoring: preliminaries



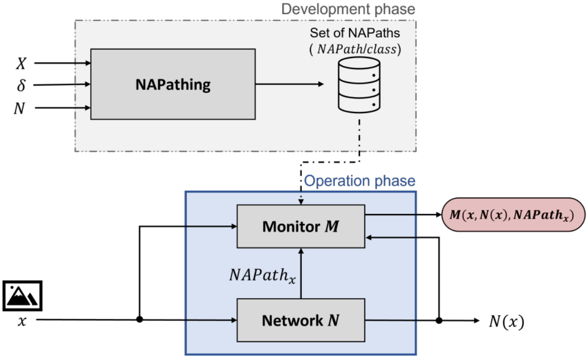
Active/inactive paths

- Active neuron (positive)
- Inactive neuron (negative)
- Path from x to y

Image classification objective

- During training
- Most common paths in the images of a same class
  $\rightarrow$ Active/inactive paths associated with this class

Monitoring architecture



Neural network

- Image x
- Classification $N(x)$

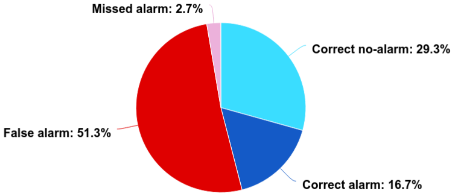
Monitor to confirm (or not) the obtained classification

- Active paths by image x
- **Alarm** if inconsistent with paths learned for the class $N(x)$
  - $\rightarrow$ Manual verification or use redundancy
  - $\rightarrow$ **Safety-critical systems** : avoid false negatives

Monitoring: tests

Weather detection: fog, rain, snow, sun

- Dataset: 1963 pictures
- ReLU network: 150528 inputs
  2 hidden layers (width 224, 84)
  4 outputs
- Large number of (false) alarms
- **Low number of missed alarms**

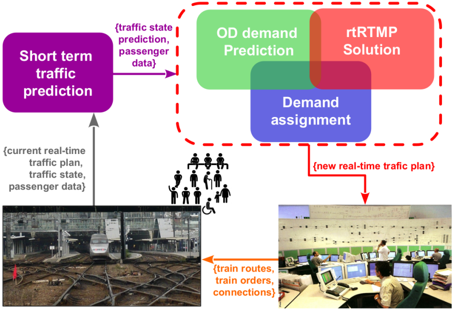


Outline

AI for railway traffic

Project SortedMobilty

- real-time Railway Traffic Management Problem: re-routing and re-scheduling to minimize delay propagation
- **ML-based short-term prediction of traffic demand** to improve **solution quality assessment in MILP**



New project ReinforceRail

- Hybridization Operational Research/AI tools for rtRTMP
- Neural networks to help reduce the size of the traffic problem

Contacts at ESTAS lab

- **Pierre-Jean Meyer**
  - Reachability analysis of neural networks
- **Fateh Boudardara**
  - Neural network reduction
  - Neural network monitoring
- **Abdelrahman Ibrahim**
  - Formal verification of neural ODE
- **Paola Pellegrini**
  - AI for railway traffic