UNIVERSITY OF OXFORD

Safety verification for deep neural networks with provable guarantees

Marta Kwiatkowska

Department of Computer Science, University of Oxford

SECML at NeurIPS 2018, 7th December 2018

Deep learning with everything

nature International weekly journal of science

Home | News | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors

Archive > Volume 542 > Issue 7639 > Letters > Article > Article metrics > News

Article metrics for:

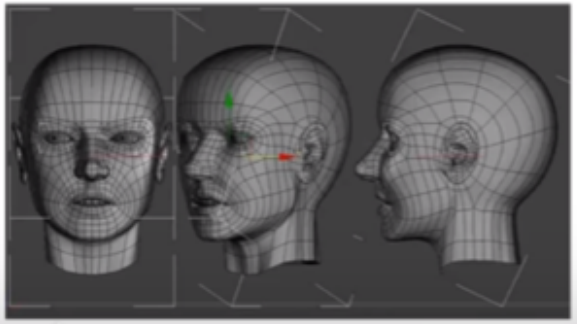
Dermatologist-level classification of skin cancer with deep neural networks

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau & Sebastian Thrun

Nature **542**, 115–118 (02 February 2017) | doi:10.1038/nature21056

Last updated: 24 July 2017 10:10:28 EDT

DeepFace
Closing the Gap to Human-Level Performance in Face Verification



Yaniv Taigman
Ming Yang
Marc'Aurelio Ranzato
Lior Wolf
- 2014

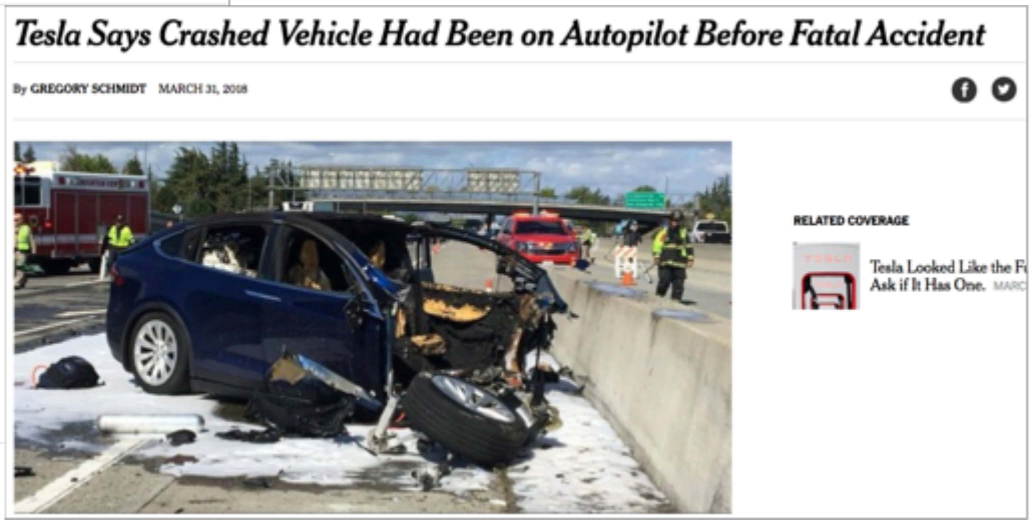
97.35% accuracy
Trained on the largest facial dataset – 4M facial images belonging to more than 4,000 identities.

Unwelcome news recently…

Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam

Leer en español

By DAISUKE WAKABAYASHI MARCH 19, 2018



Tesla Says Crashed Vehicle Had Been on Autopilot Before Fatal Accident

By GREGORY SCHMIDT MARCH 31, 2018

RELATED COVERAGE

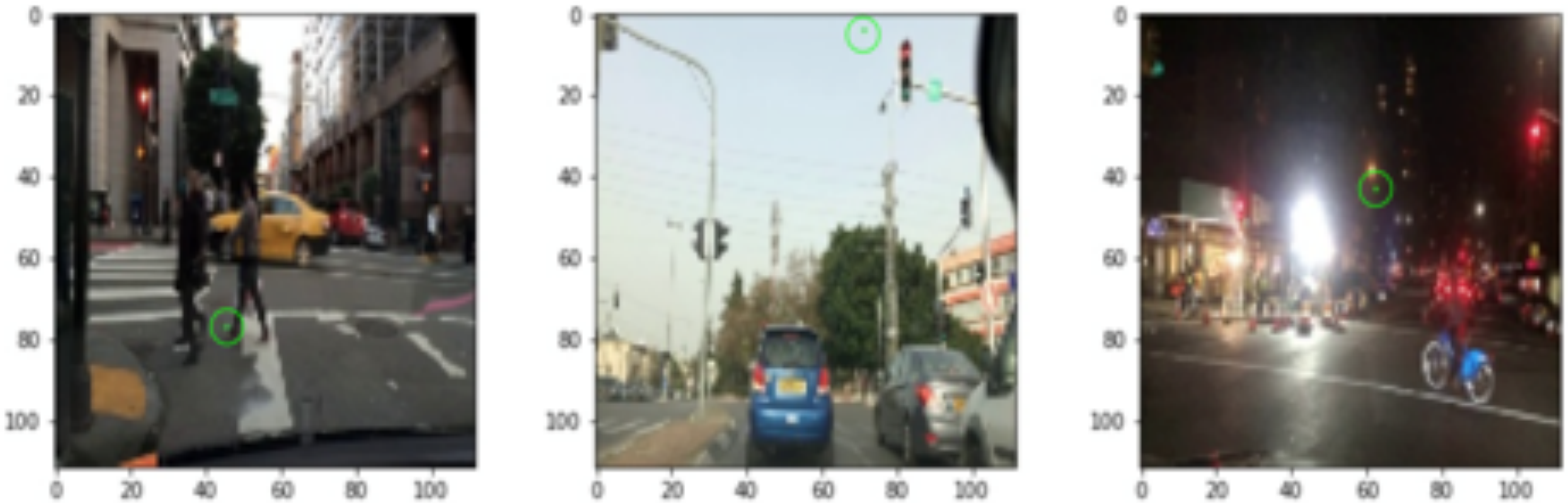
Fatal Tesla Crash Raises New Questions About Autopilot System

U.S. Safety Agency Criticizes Tesla Crash Data Release

How can this happen if we have 99.9% accuracy?

https://www.youtube.com/watch?v=B2pDFjIvrIU

Should we worry about safety?



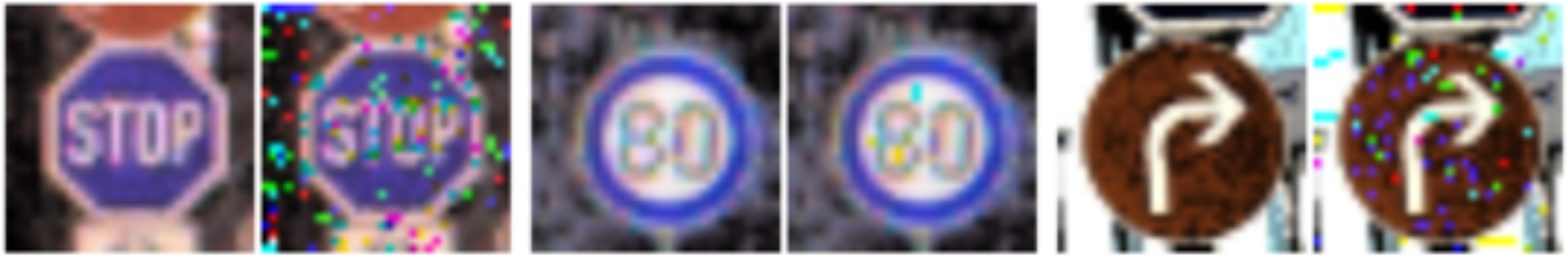
(a) (b) (c)

Red light classified as green with (a) 68%, (b) 95%, (c) 78% confidence after one pixel change.

- TACAS 2018, https://arxiv.org/abs/1710.07859

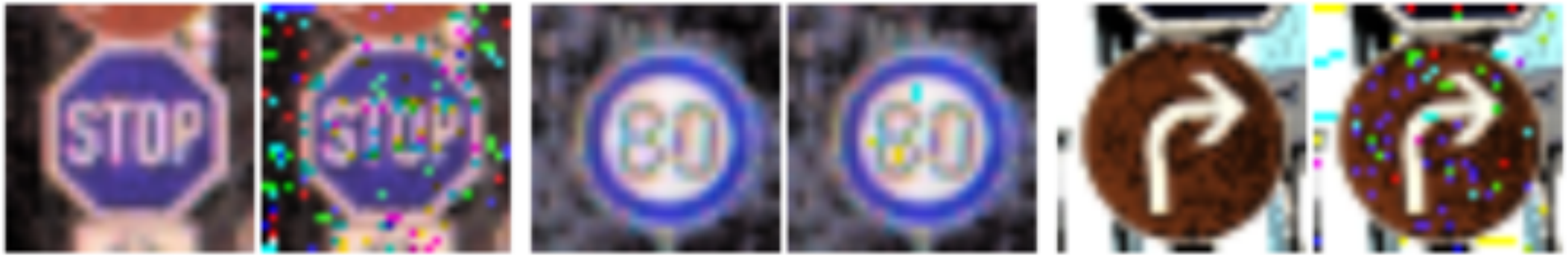
Can we verify that such behaviour cannot occur?

German traffic sign benchmark…



stop	30m speed limit	80m speed limit	30m speed limit	go right	go straight

German traffic sign benchmark…



stop	30m speed limit	80m speed limit	30m speed limit	go right	go straight

Confidence 0.999964 0.99

Aren’t these artificial?

Real traffic signs in Alaska!

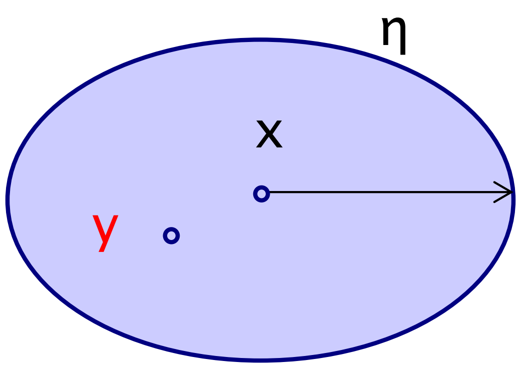
Need to consider physical attacks, not only digital…

This talk

- Progress in developing methodology to provide provable guarantees of safety of classification decisions
- Focus on local robustness against adversarial manipulations
- Automated verification
  - search/SMT: CAV 2017, https://arxiv.org/abs/1610.06940
  - game: TACAS 2018, https://arxiv.org/abs/1710.07859
- Reachability analysis
  - global optim: IJCAI 2018, https://arxiv.org/abs/1805.02242
- Testing with coverage guarantees
  - concolic: ASE 2018, https://arxiv.org/abs/1805.00089
- Probabilistic safety
  - Bayesian GP: AAAI 2019, https://arxiv.org/abs/1809.06452

Safety of classification decisions

- Safety assurance process is complex
- Here focus on safety at a point as part of such a process
  - same as pointwise robustness…



- Assume given
  - trained network $f : D \rightarrow \{c_1,\ldots c_k\}$
  - diameter for support region η
  - norm, e.g. L^2, L^∞
- Define safety as invariance of classification decision
  - i.e. $\nexists y \in \eta$ such that $f(x) \neq f(y)$
- Also wrt family of safe manipulations
  - e.g. scratches, weather conditions, camera angle, etc

Safety verification

- Automated verification (= ruling out adversarial examples)
  - discretise the region, exhaustively search for misclassifications
  - provable guarantee of decision safety if adv. example not found
  - (assumptions needed to ensure finiteness of search)
- The approach
  - reduction to linear arithmetic (counting problem), use SMT
  - propagate verification layer by layer
- This differs from heuristic search for adversarial examples
  - no guarantee of precise adversarial examples
  - no guarantee of exhaustive search even if we iterate

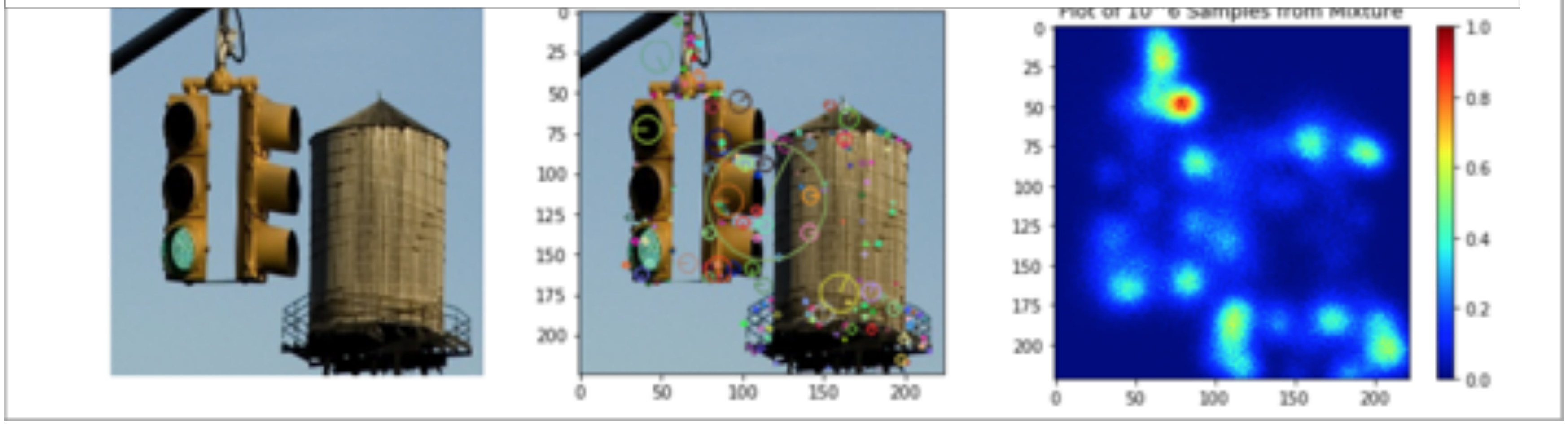
- But scalability remains an issues, employ various heuristics…

- CAV 2017, https://arxiv.org/abs/1610.06940

Feature-based representation

- Employ the SIFT algorithm to extract features
- Reduce dimensionality by focusing on salient features
- Use a Gaussian mixture model in order to assign each pixel a probability based on its perceived saliency

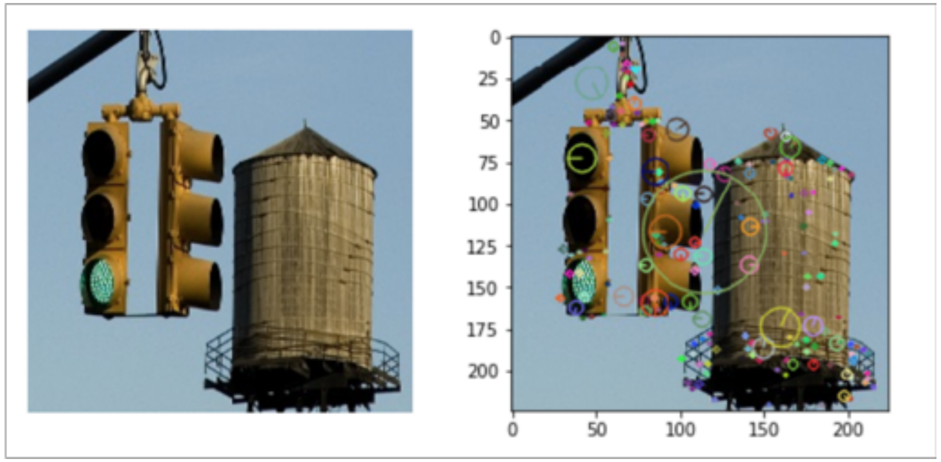
$$\mathcal{G}_{i,x} = \frac{1}{\sqrt{2\pi\lambda_{i,s}^2}} exp\left(\frac{-(p_x - \lambda_{i,x})^2}{2\lambda_{i,s}^2}\right) \qquad \mathcal{G}_{i,y} = \frac{1}{\sqrt{2\pi\lambda_{i,s}^2}} exp\left(\frac{-(p_y - \lambda_{i,y})^2}{2\lambda_{i,s}^2}\right)$$



TACAS 2018, https://arxiv.org/abs/1710.07859

Game-based search

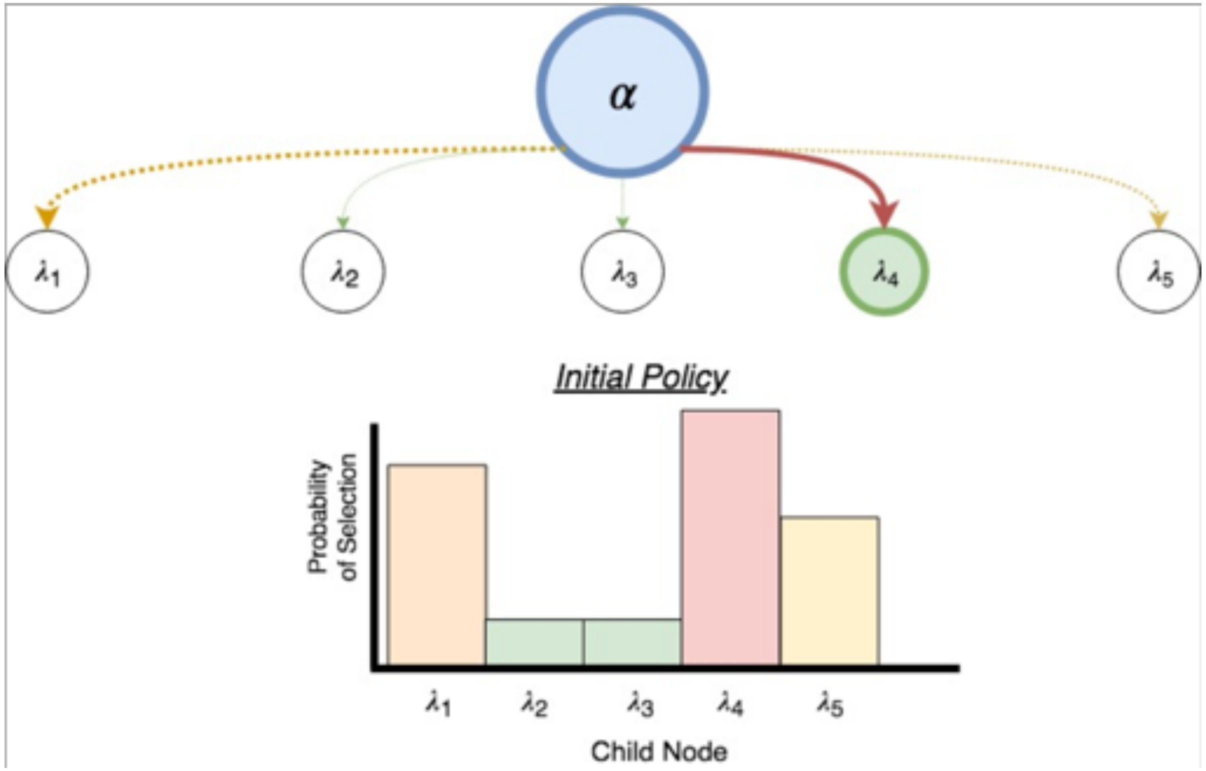
- Goal is finding adv. example, reward inverse of distance
- Player 1 selects the feature that we will manipulate



- Each feature represents a possible move for player 1
- Player 2 then selects the pixels in the feature to manipulate
- Use Monte Carlo tree search to explore the game tree, while querying the network to align features
- Method black/grey box, can approximate the maximum safe radius for a given input

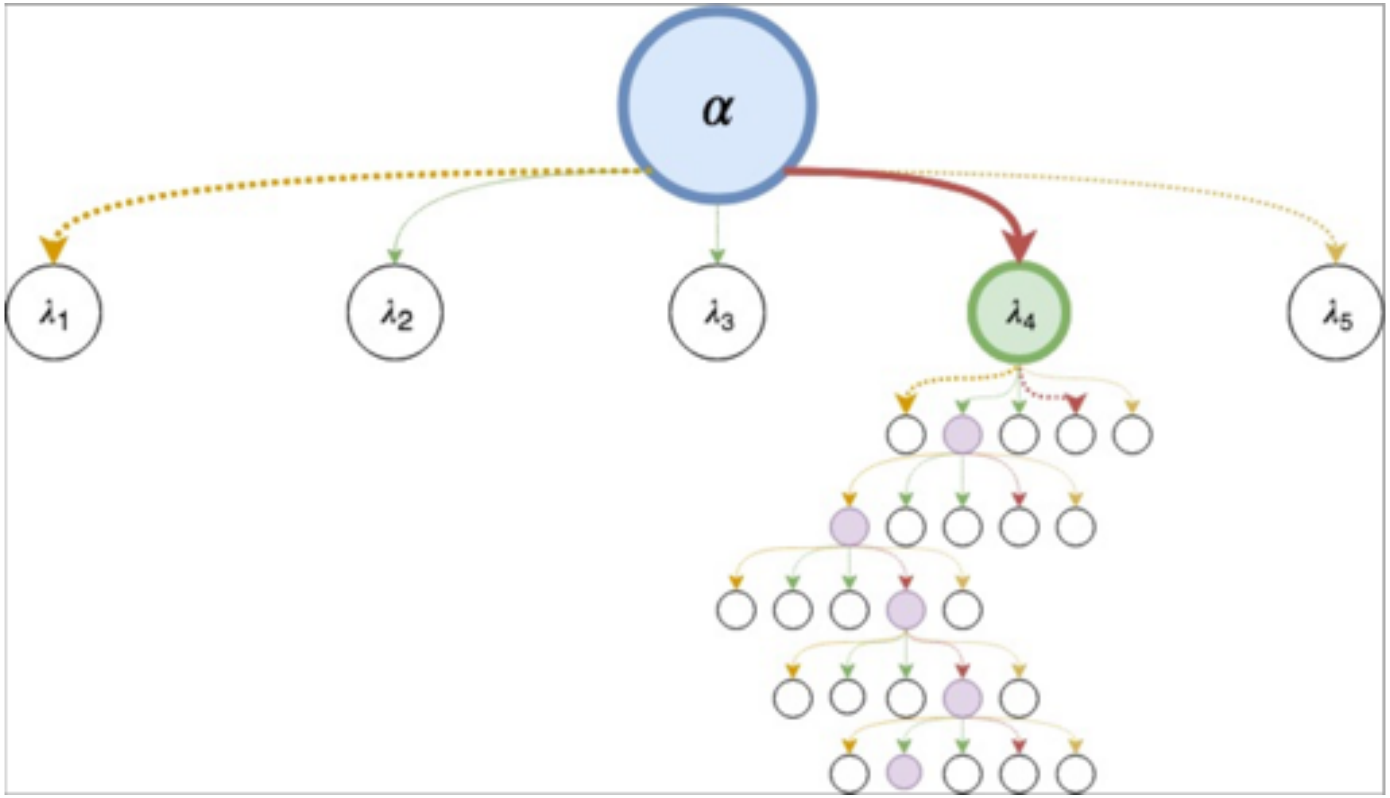
MCTS: selection/expansion

- The root of the tree represents the original image, and each child represents a potential manipulated image
- First, select a manipulation based on each player's strategy
- If the child has never been selected from previously then we "expand" the tree to select a new leaf.



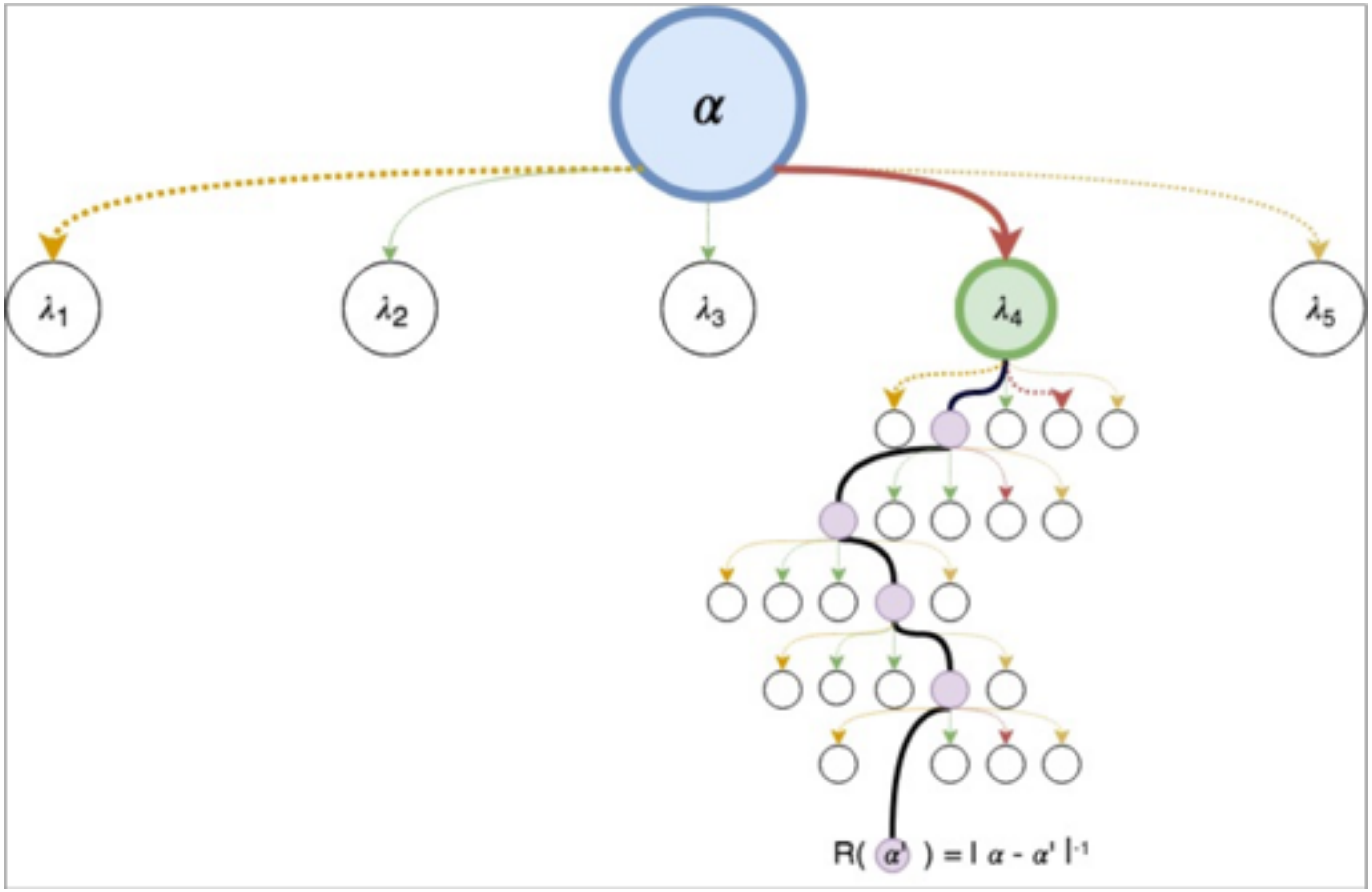
MCTS: simulation

- After a new child has been added to the tree, we approximate the reward of visiting this child by continuously searching the tree until we have either timed out or hit an adversarial example
- These nodes are not recorded as a part of the partial tree

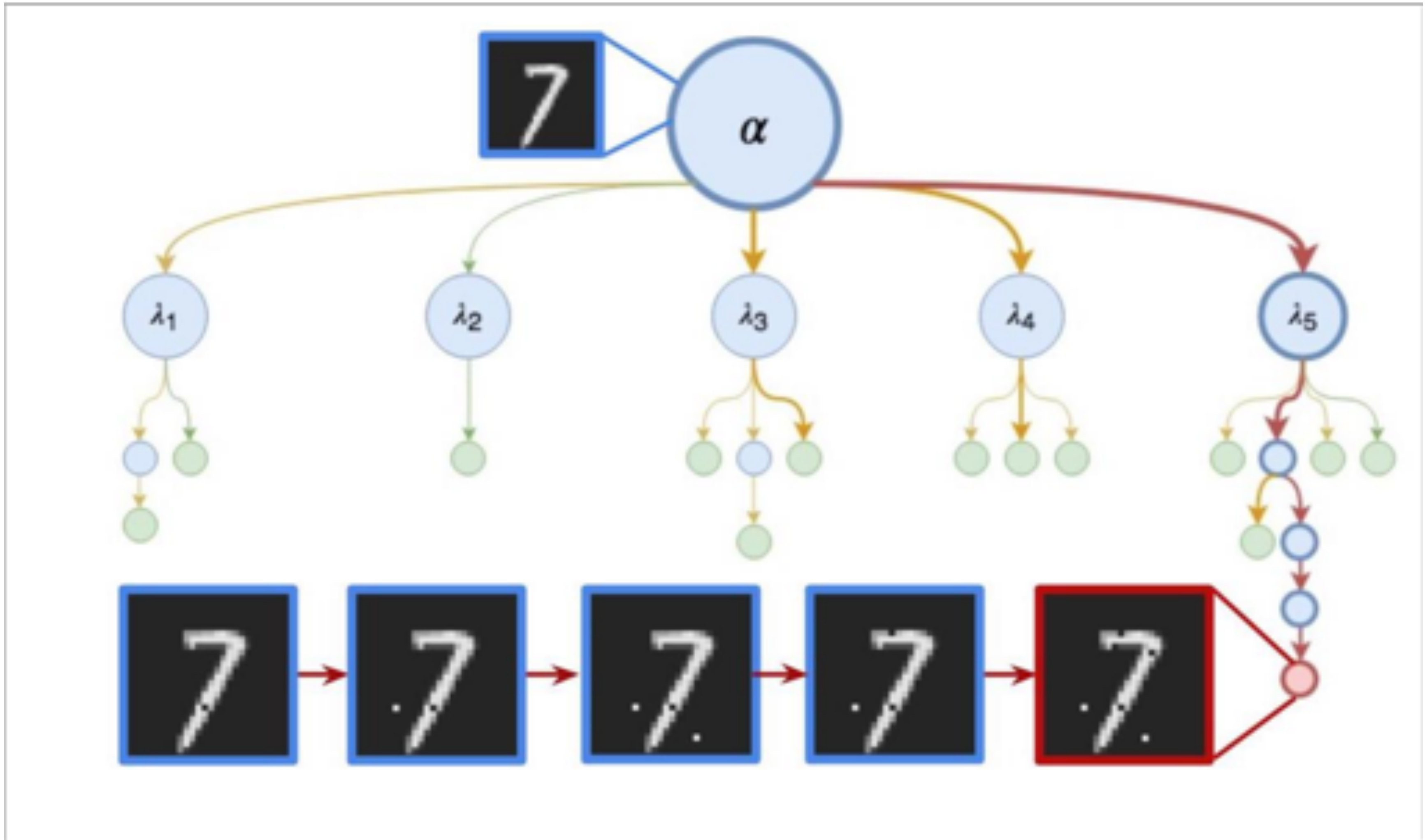


MCTS: backpropagation

- After we have terminated the tree, we calculate the reward, and backpropagate that reward up the tree to update our exploration policy (update each player's strategies)

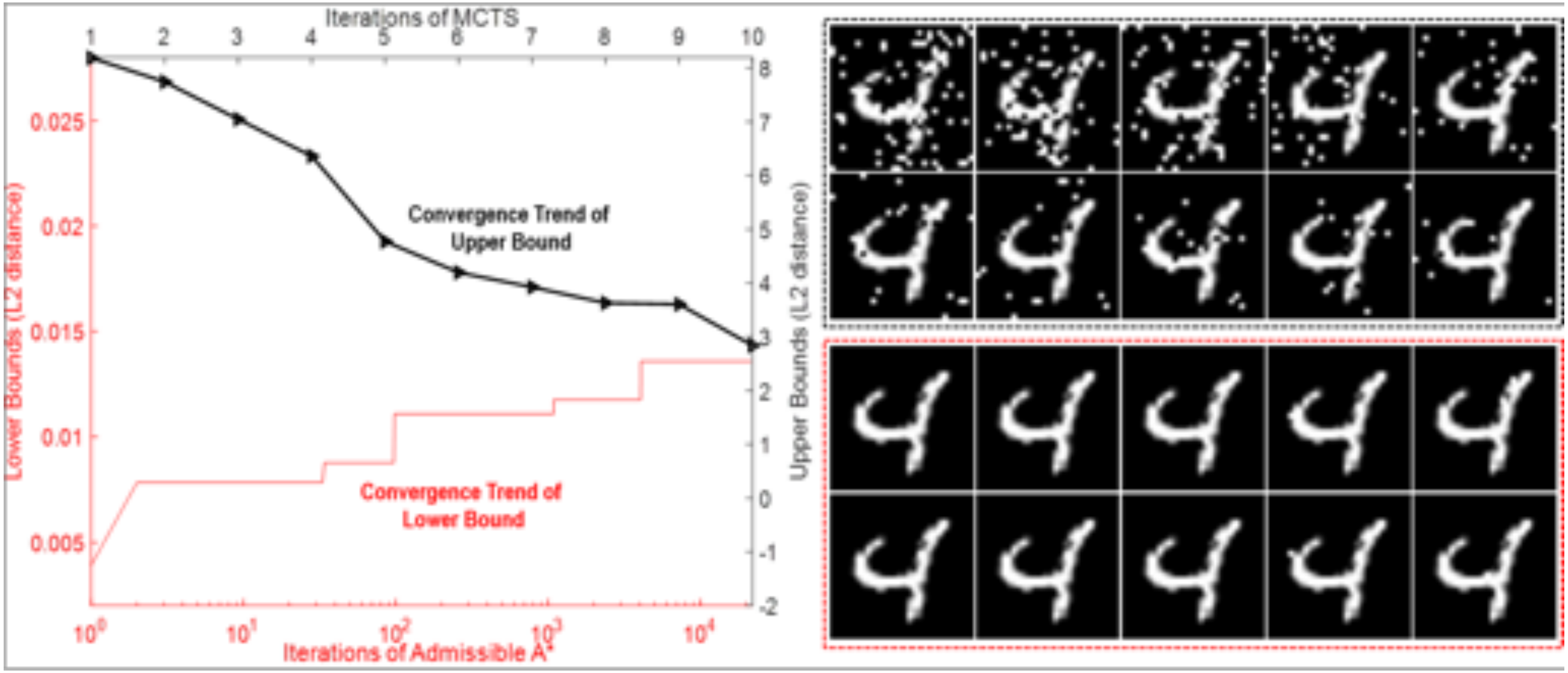


Tree expands until example is found



Now also lower bounds (MNIST)

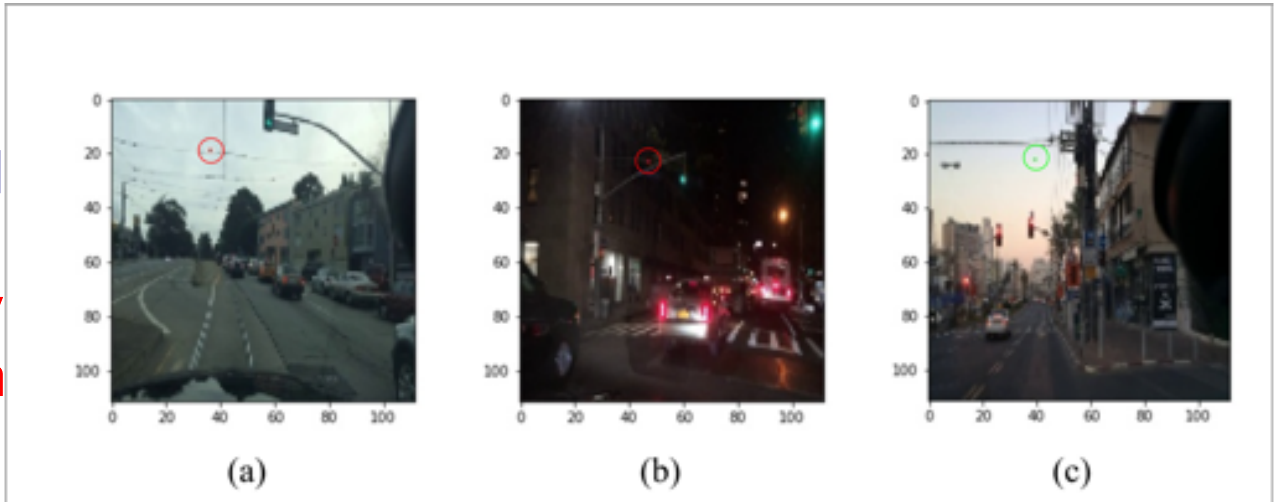
- Convergence of lower and upper bounds on maximum safe radius



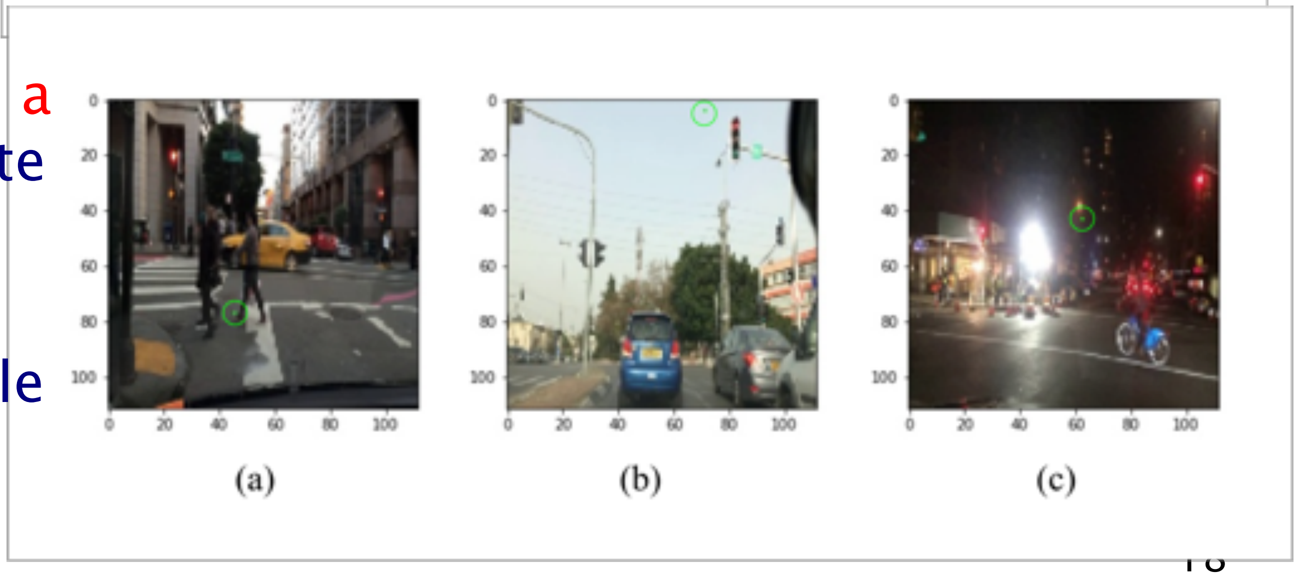
- See arXiv:1807.0357

Evaluating safety-critical scenarios: Nexar

- Using our Game-based Monte Carlo Tree Search method we were able to reduce the accuracy of the network form 95% to 0%
- On average, each input took less than a second to manipulate (.304 seconds)
- On average each image was vulnerable to 3 pixel changes



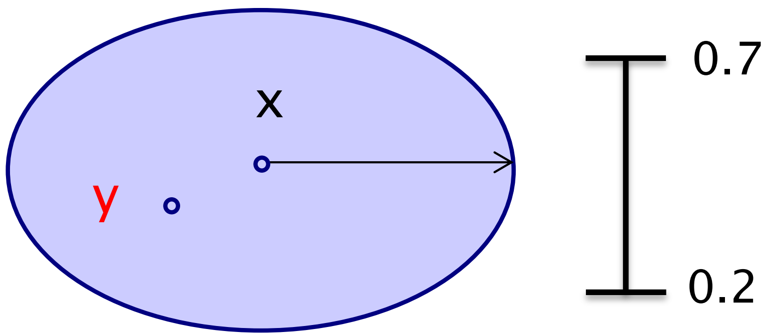
(a) (b) (c)



(a) (b) (c)

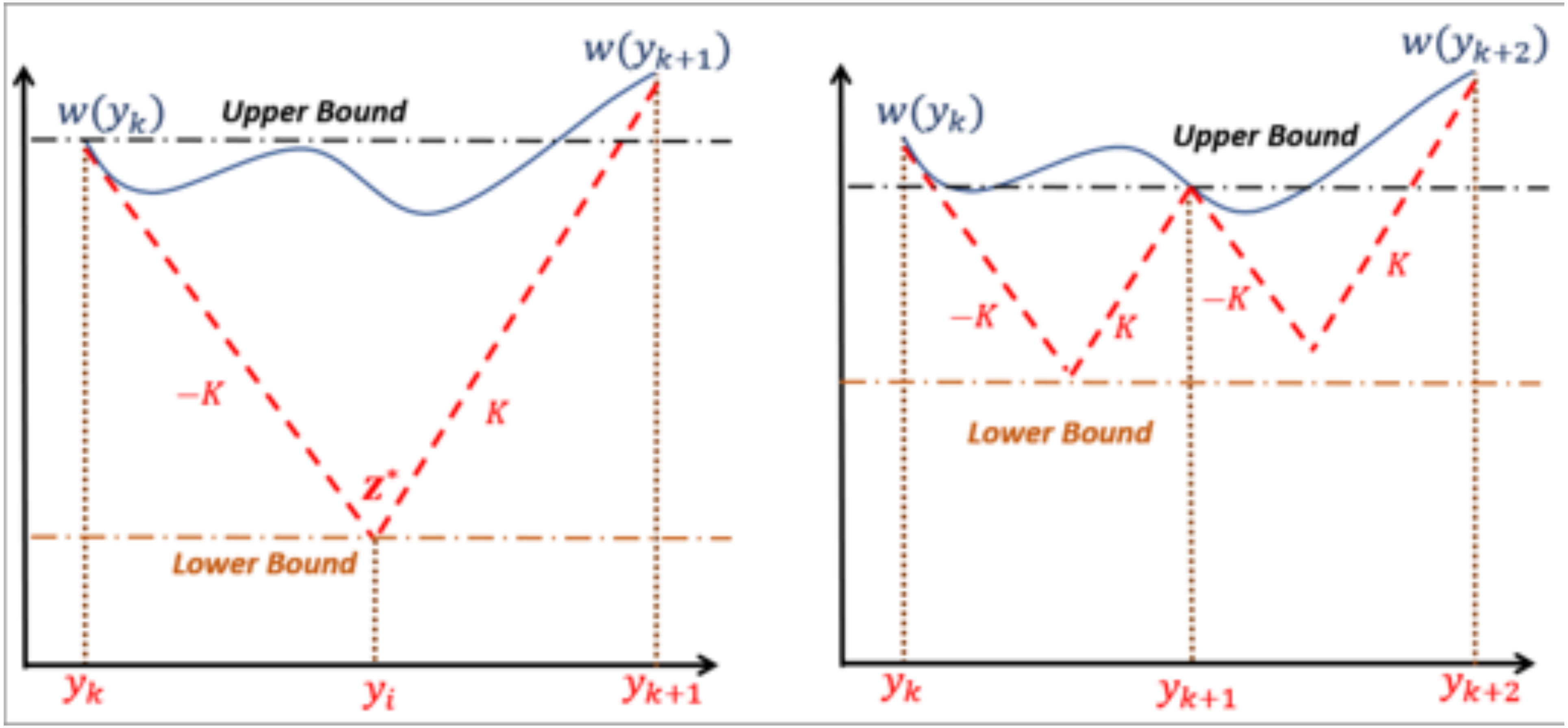
Alternative approach: reachability analysis

- Rather than search the discretized region, can we compute the reachable values?
- Under assumption of Lipschitz continuity
  - for $x \in \eta$, compute maximum/minimum value of $f(\eta)$
  - using global optimisation
  - anytime fashion
- Gives provable guarantees
  - best/worst case confidence values
  - pointwise confidence diameter
  - can average over input distribution
- Method NP-complete
  - wrt the number of input dimensions, not number of neurons



- IJCAI 2018, https://arxiv.org/abs/1805.02242

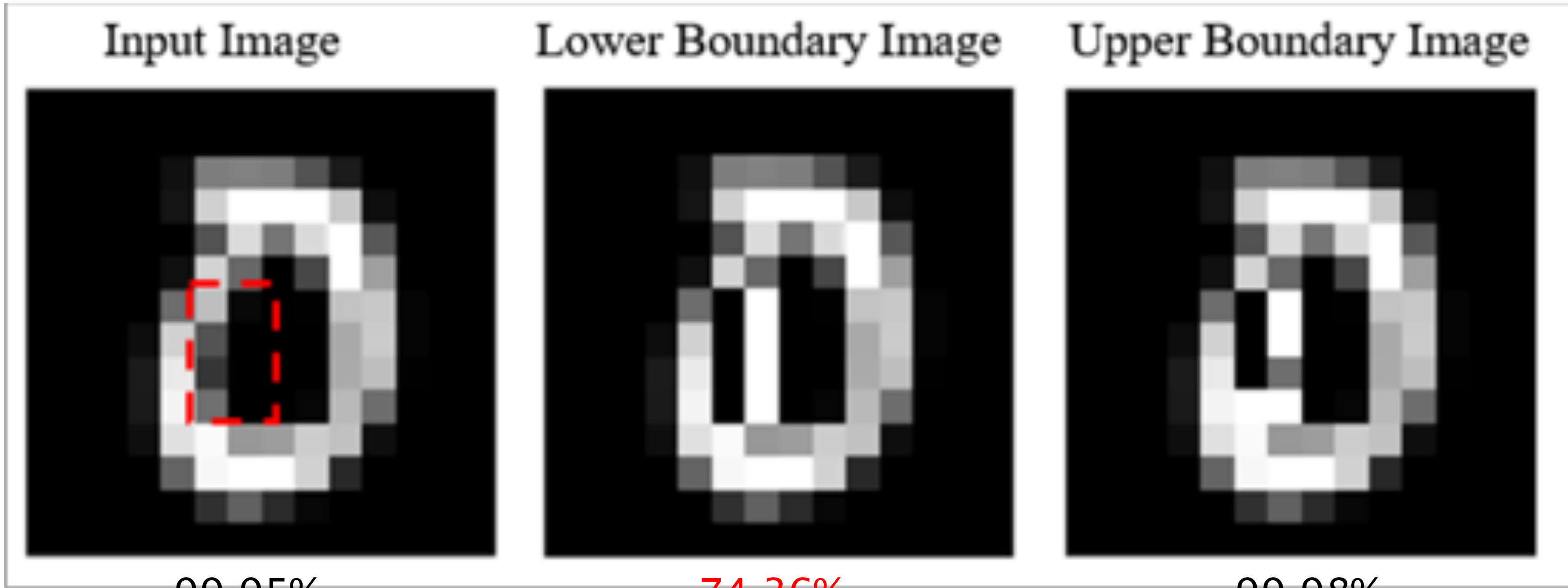
Global optimization: main idea



- Adaptive nested optimization, asymptotic convergence
  - construct a series of lower and upper bounds
- K – Lipschitz constant

MNIST example

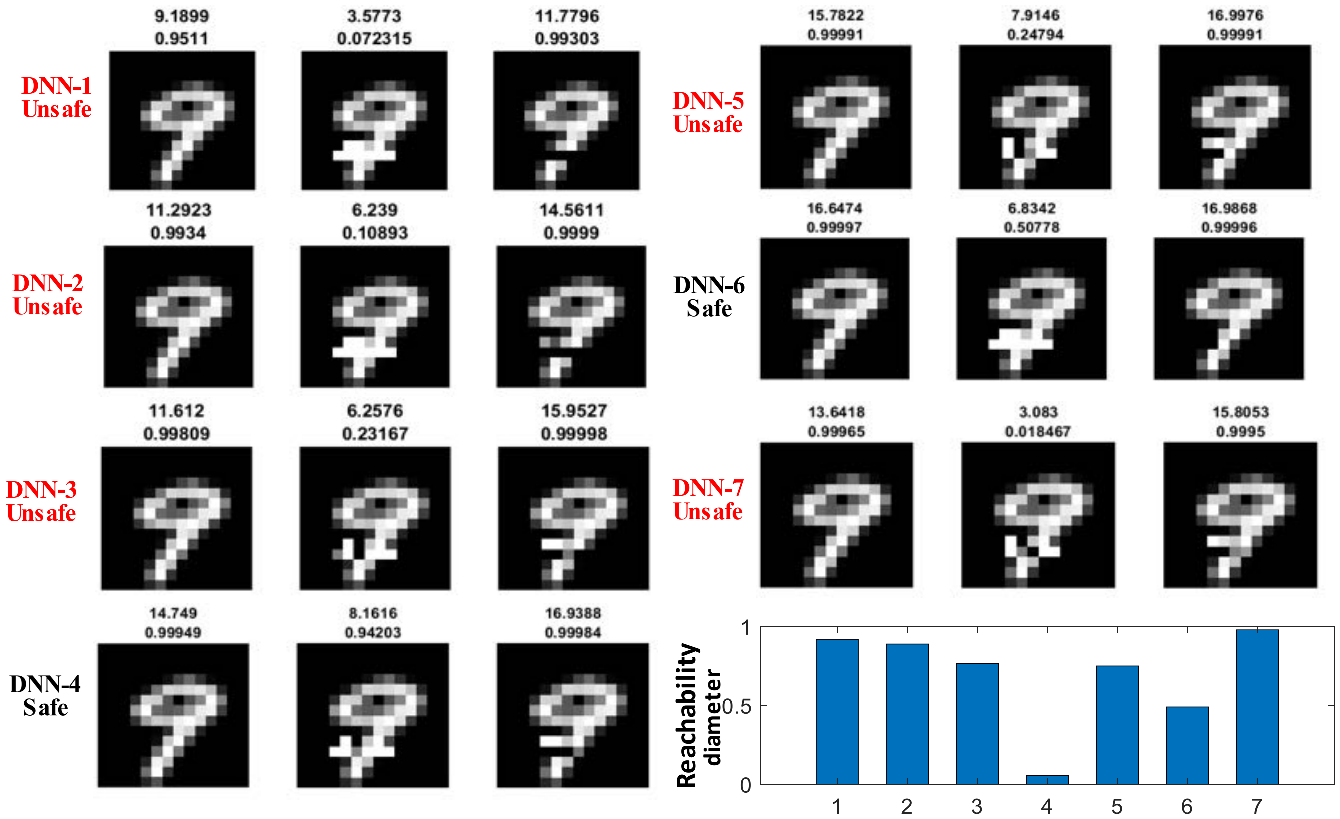
- Take an image and select a feature within it



99.95% confidence	74.36% lower bound	99.98% upper bound

- Safety verification for the feature
  - manipulating the feature can only reduce confidence to 74.36%

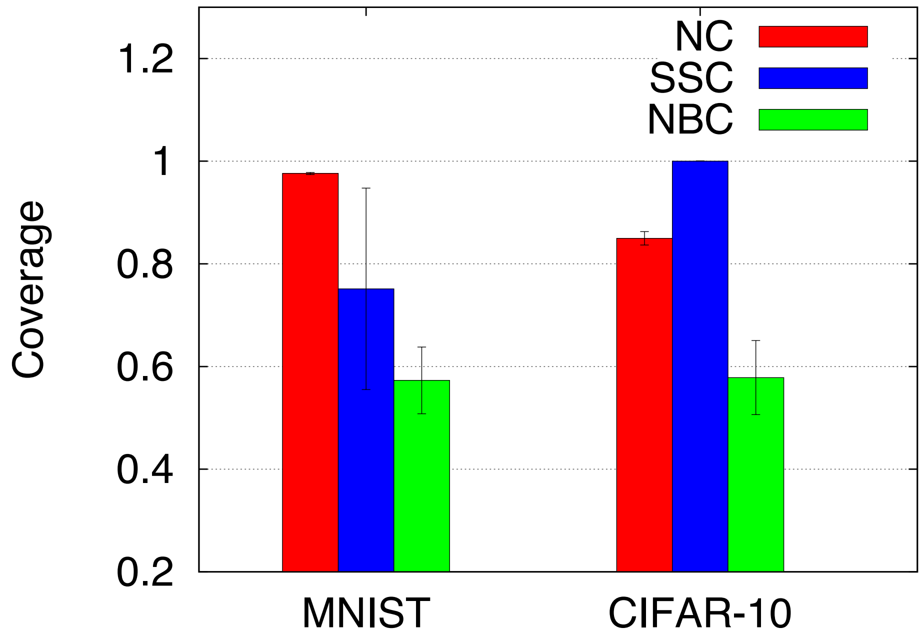
MNIST network comparison



- Showing pointwise confidence diameter
- Can obtain global robustness evaluation by averaging wrt the test data distribution

Safety testing, guaranteed coverage

- Often provable guarantees on network outputs beyond reach
- Concolic testing (concrete+symbolic)
  - for test goal, generate test cases
  - alternating between concrete execution & symbolic analysis
- Test coverage criteria
  - specified in quantified linear arithmetic over rationals
- Range of coverage metrics
  - neuron coverage (NC), neuron boundary coverage (NBC), modified condition/decision (MC/DC), Lipschitz continuity, etc

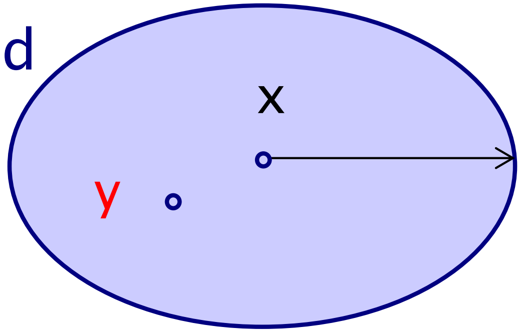


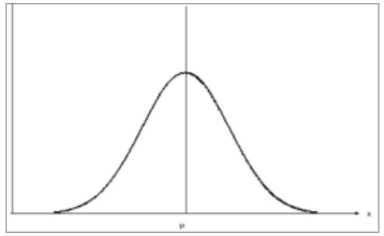
ASE 2018, https://arxiv.org/abs/1805.00089

Probabilistic guarantees

- Requiring that no adversarial examples exist too strict
- Need to probabilistic guarantees: probability that local perturbations result in predictions that are close to original

- Work with Bayesian inference and
- Gaussian processes
- Define safety with prob $1-\varepsilon$





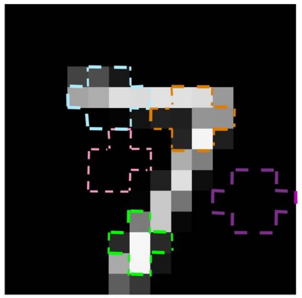
$$Prob(\exists y \in \eta \text{ s.t. } ||f(x)-f(y)||>\delta \mid D) \leq \varepsilon$$

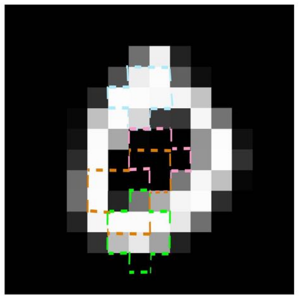
- i.e. conditioned on training data D
- NB differs from pointwise thresholding in Bayesian deep learning

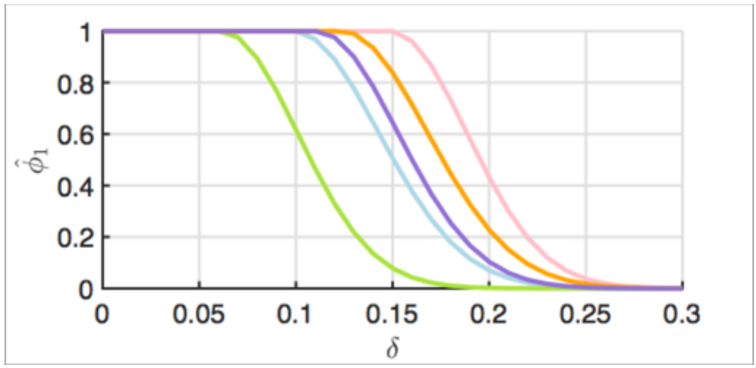
AAAI 2019, https://arxiv.org/abs/1809.06452

Probabilistic guarantees for NNs

- Computation for general stochastic processes intractable
- For GPs, can obtain tight upper bounds by
  - approximating extrema of mean and variance for a test point
  - using Borell-TIS inequality
  - and solving optimization problems (analytical or convex opt)
- Applies to fully-connected (and convolutional) neural networks in the limit of infinitely many neurons…







- Scalability continues to be an issue for NNs

Conclusion

- Deep learning should be more critically evaluated when put into practice in safety- and security-critical situations
- Adversarial examples help in understanding the robustness of DNN decision boundaries
- Overviewed methods for safety verification/testing of deep neural networks
  - search-based and feature-guided exploration, with guarantees
  - reachability computation for Lipschitz continuous networks
  - test coverage guarantees
  - probabilistic guarantees in a Bayesian framework
- Future work
  - how best to use adversarial examples: training vs logic
  - scalability for probabilistic guarantees
  - more complex properties

Acknowledgements

- My group and collaborators in this work
- Project funding
  - ERC Advanced Grant
  - EPSRC Mobile Autonomy Programme Grant

- See also
  - VERIWARE www.veriware.org
  - PRISM www.prismmodelchecker.org