VeTSS, Cambridge

25th September 2019

Safety verification for deep neural networks with provable guarantees

Prof. Marta Kwiatkowska

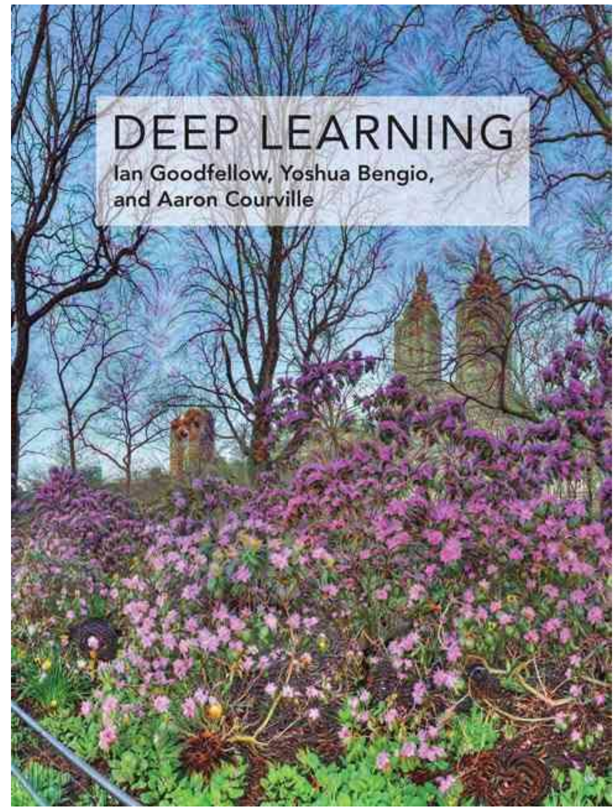
Department of Computer Science
University of Oxford

The unstoppable rise of deep learning

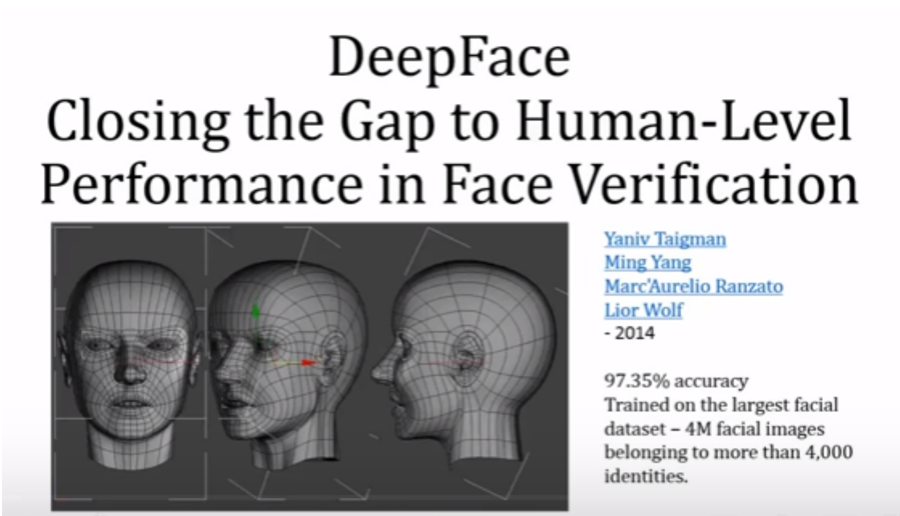
- Neural networks timeline

  1940s First proposed
  1998 Convolutional nets
  2006 Deep nets trained
  2011 Rectifier units
  2015 Vision breakthrough
  2016 Win at Go
  2019 Turing Award

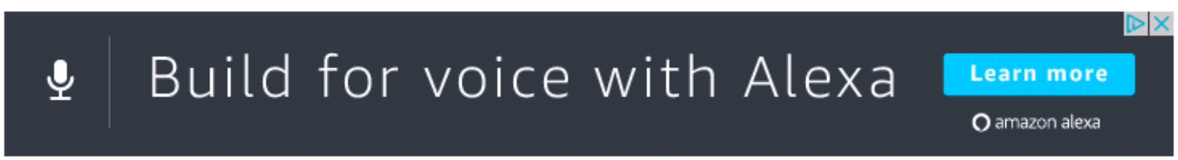
- Enabled by
  - Big data
  - Flexible, easy to build models
  - Availability of GPUs
  - Efficient inference



Much interest from tech companies,



Google Translate—here shown on a mobile phone—will use deep learning to improve its translations between texts.



...healthcare,

nature International weekly journal of science

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

The Stanford University team said the findings were "incredibly exciting" and would now be tested in clinics.

Eventually, they believe using AI could revolutionise healthcare by turning anyone's smartphone into a cancer scanner.

Cancer Research UK said it could become a useful tool for doctors.

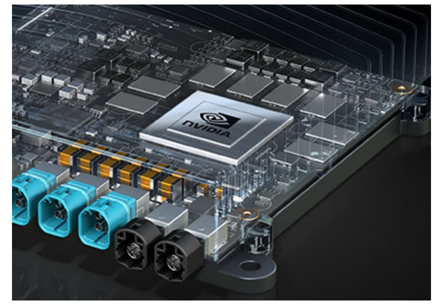
The AI was repurposed from software developed by Google that had learned to spot the difference **between images of cats and dogs**.

…and automotive industry

https://www.youtube.com/watch?v=mCmO_5ZxdvE

What you have seen

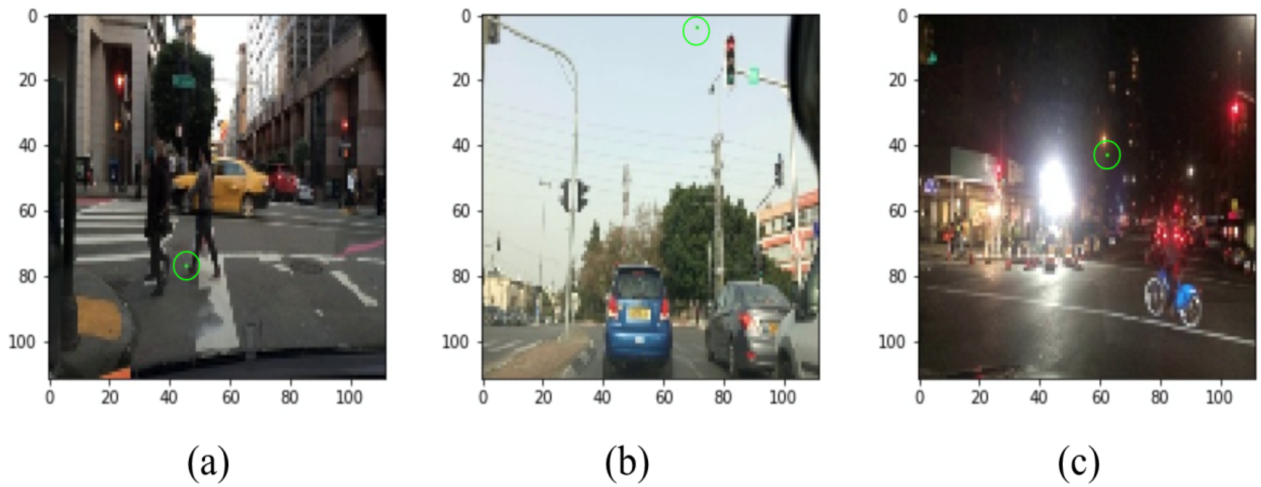
- PilotNet by NVIDIA (regression problem)
  - end-to-end controller for self-driving cars
  - neural network
  - lane keeping and changing
  - trained on data from human driven cars
  - runs on DRIVE PX 2



- Traffic sign recognition (classification problem)
  - conventional object recognition
  - neural network solutions already planned…
- BUT
  - neural networks don’t come with rigorous guarantees!

PilotNet https://arxiv.org/abs/1604.07316

Should we worry about safety of self-driving?



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?

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

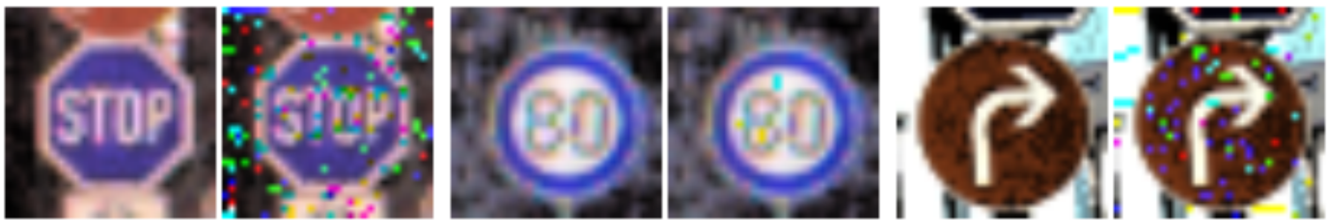
Tesla Looked Like the Fu
Ask if It Has One. MARC

Fatal Tesla Crash Raises New Questions About Autopilot System

U.S. Safety Agency Criticizes Tesla Crash Data Release

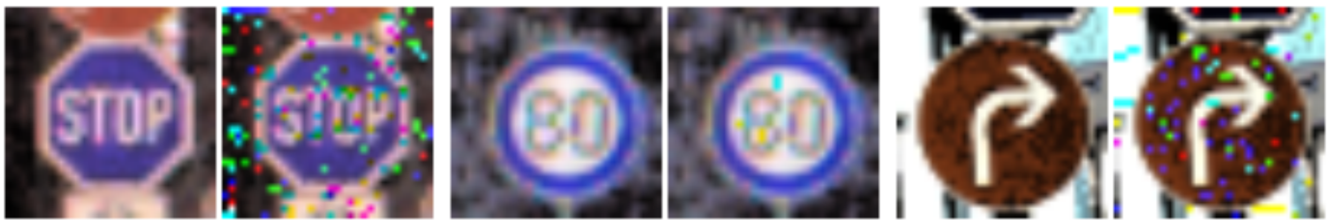
How can this happen if we have 99.9% accuracy?

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

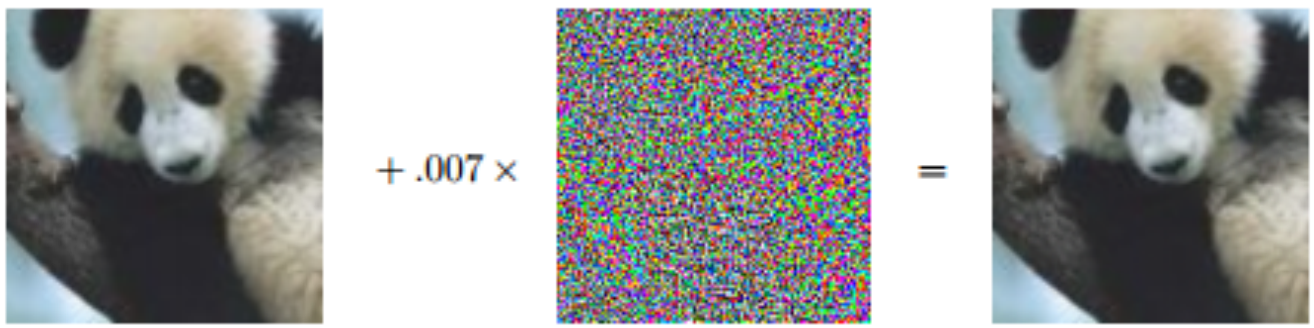
Confidence 0.999964 0.99

Aren't these artificial?

Real traffic signs in Alaska!

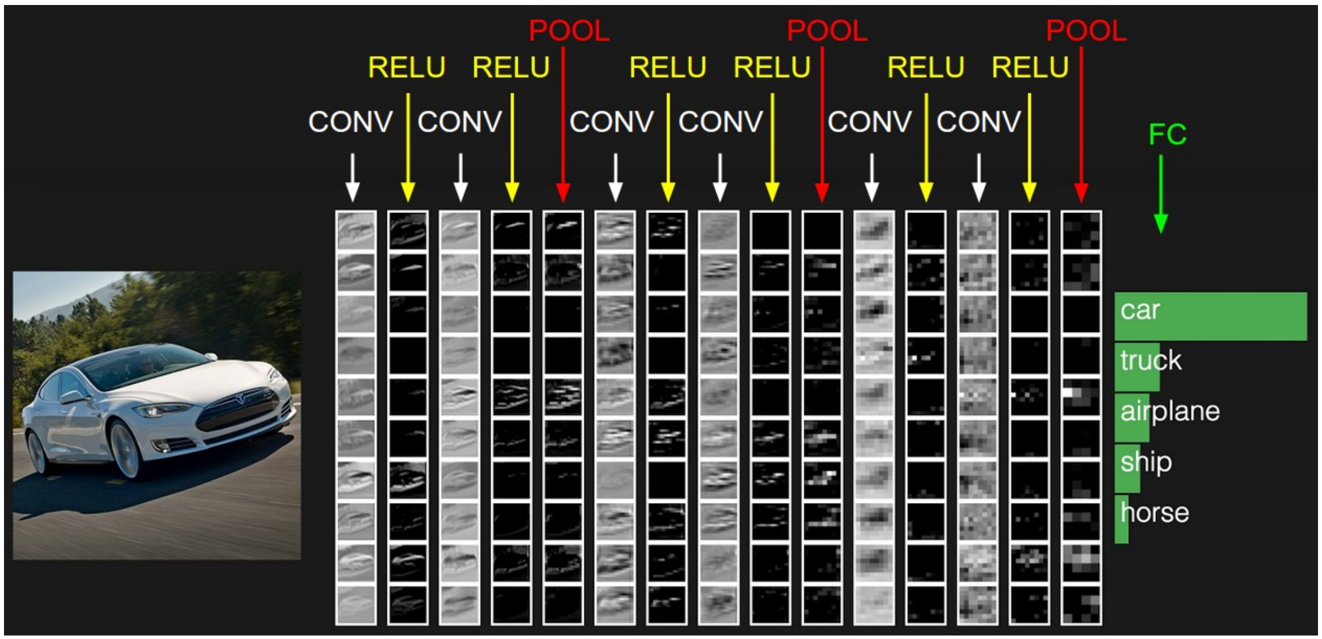
Must not overfocus on digital attacks! Need to consider physical attacks...

Deep neural networks can be fooled!



- They are unstable wrt adversarial perturbations
  - often imperceptible changes to the image [Szegedy et al 2014, Biggio et al 2013 ...]
  - sometimes artificial white noise
  - practical attacks, potential security risk
  - transferable between different architectures
  - not just image classification: also images segmentation, pose recognition, sentiment analysis...

Deep feed–forward neural network



Convolutional multi–layer network

http://cs231n.github.io/convolutional-networks/#conv

Problem setting

- Assume
  - vector spaces $D=D_{L0}, D_{L1}, \ldots, D_{Ln}$, one for each layer
  - $F : D \rightarrow \{c_1,\ldots c_k\}$ classifier function, e.g. modelling human perception ability

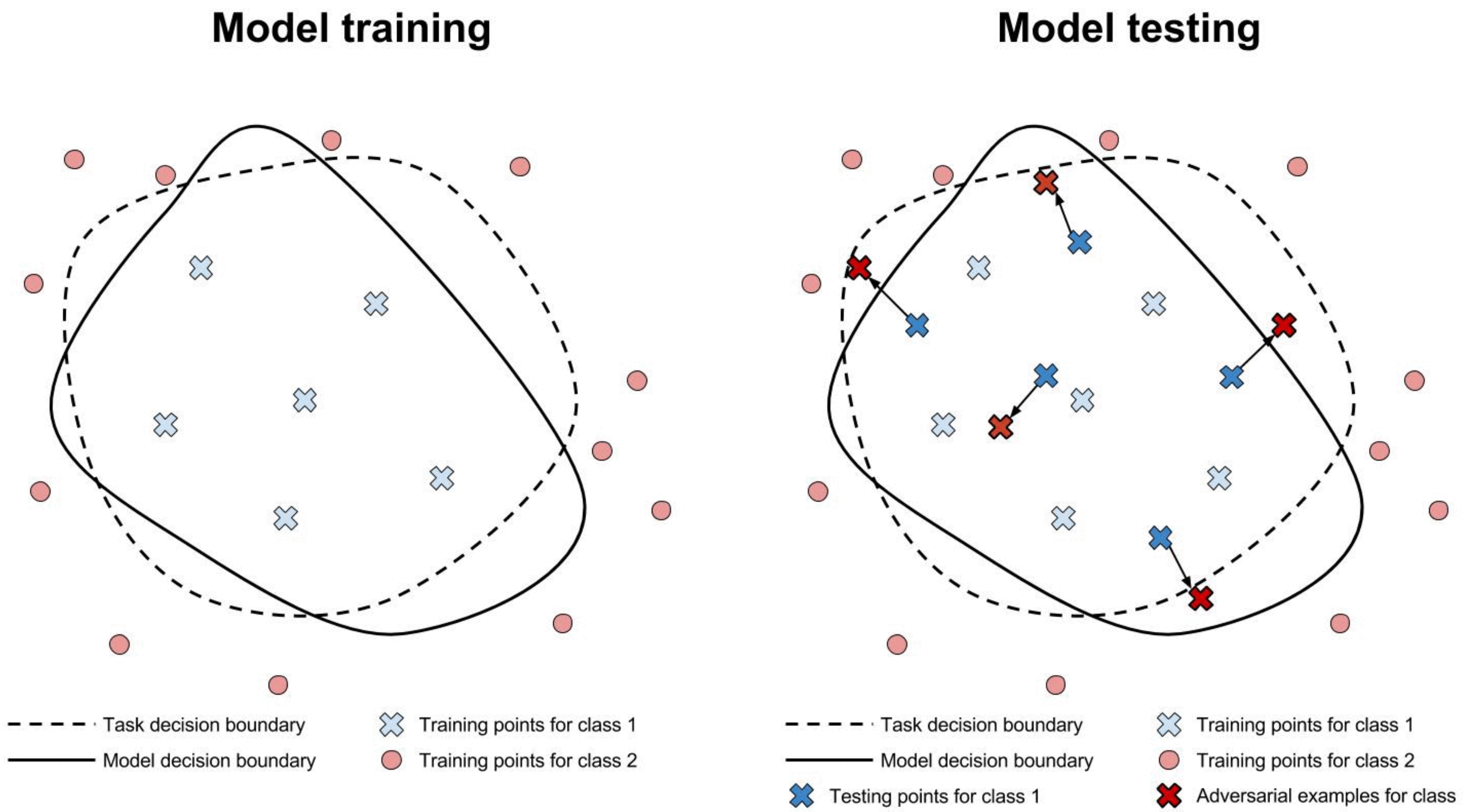
- The neural network $f : D \rightarrow \{c_1,\ldots c_k\}$ approximates F from M training examples $\{(x_i,c_i)\}_{i=1..M}$
  - built from activation functions $\phi_0, \phi_1, \ldots, \phi_n$, one for each layer
  - for point (image) $x \in D_{L0}$, its activation in layer k is

    $$\alpha_{x,k} = \phi_k(\phi_{k-1}(\ldots\phi_1(x)))$$

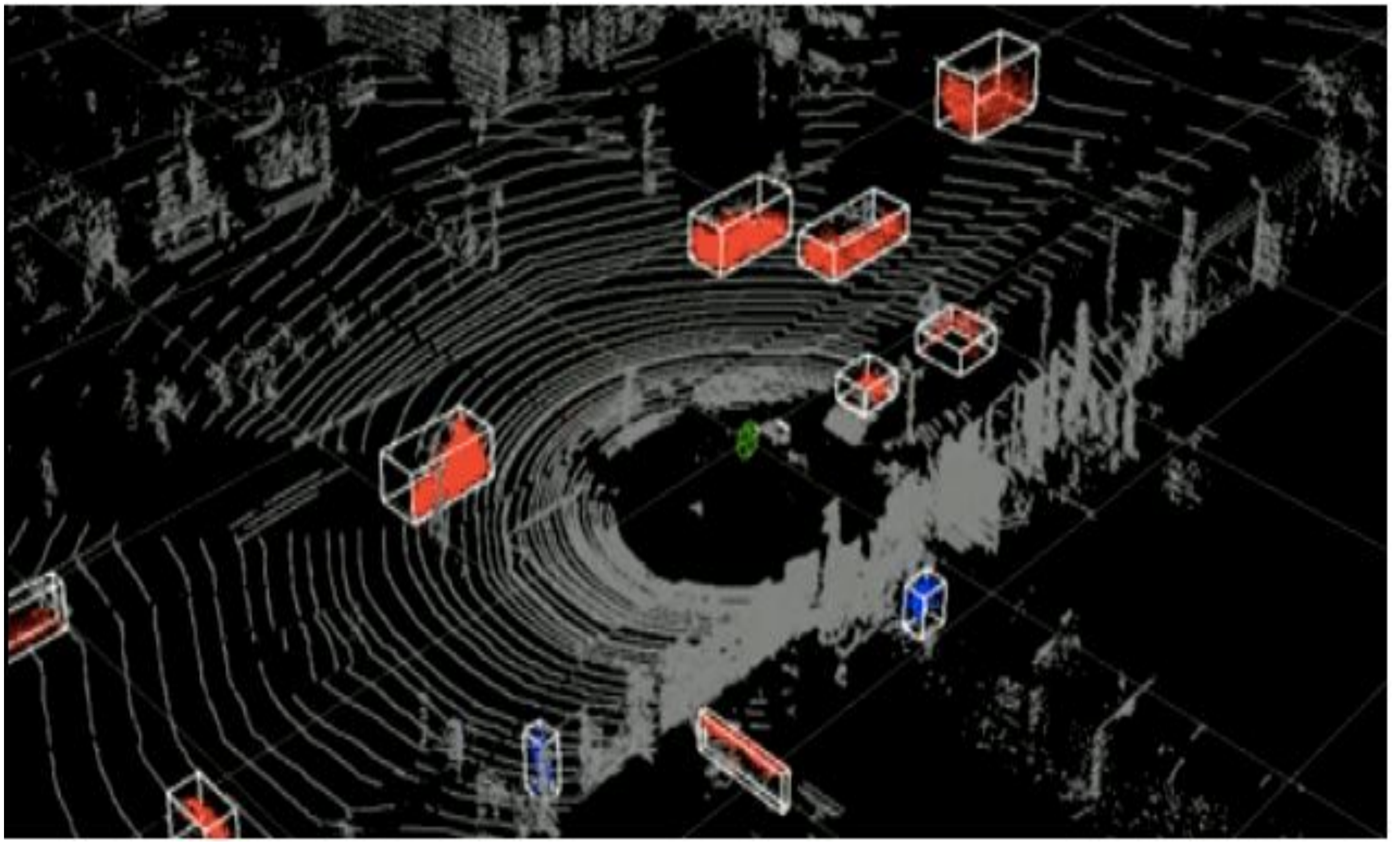
  - where $\phi_k(x) = \sigma(xW_k+b_k)$ linear transformation and $\sigma(x) = \max(x,0)$
  - W_k learnable weights, b_k bias, σ ReLU

- Notation
  - $f(x)$ is the class assigned to input x by the network

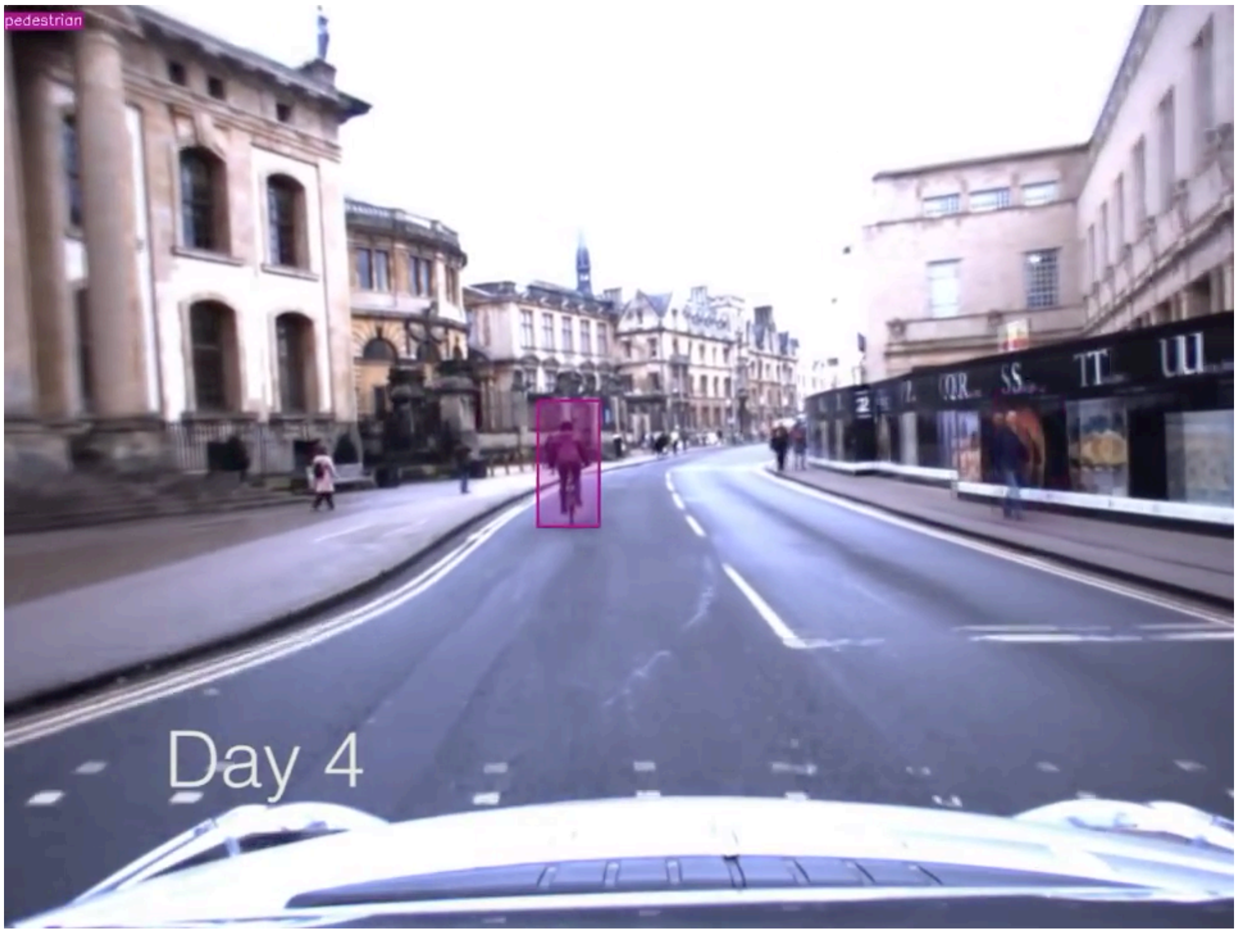
Training vs testing



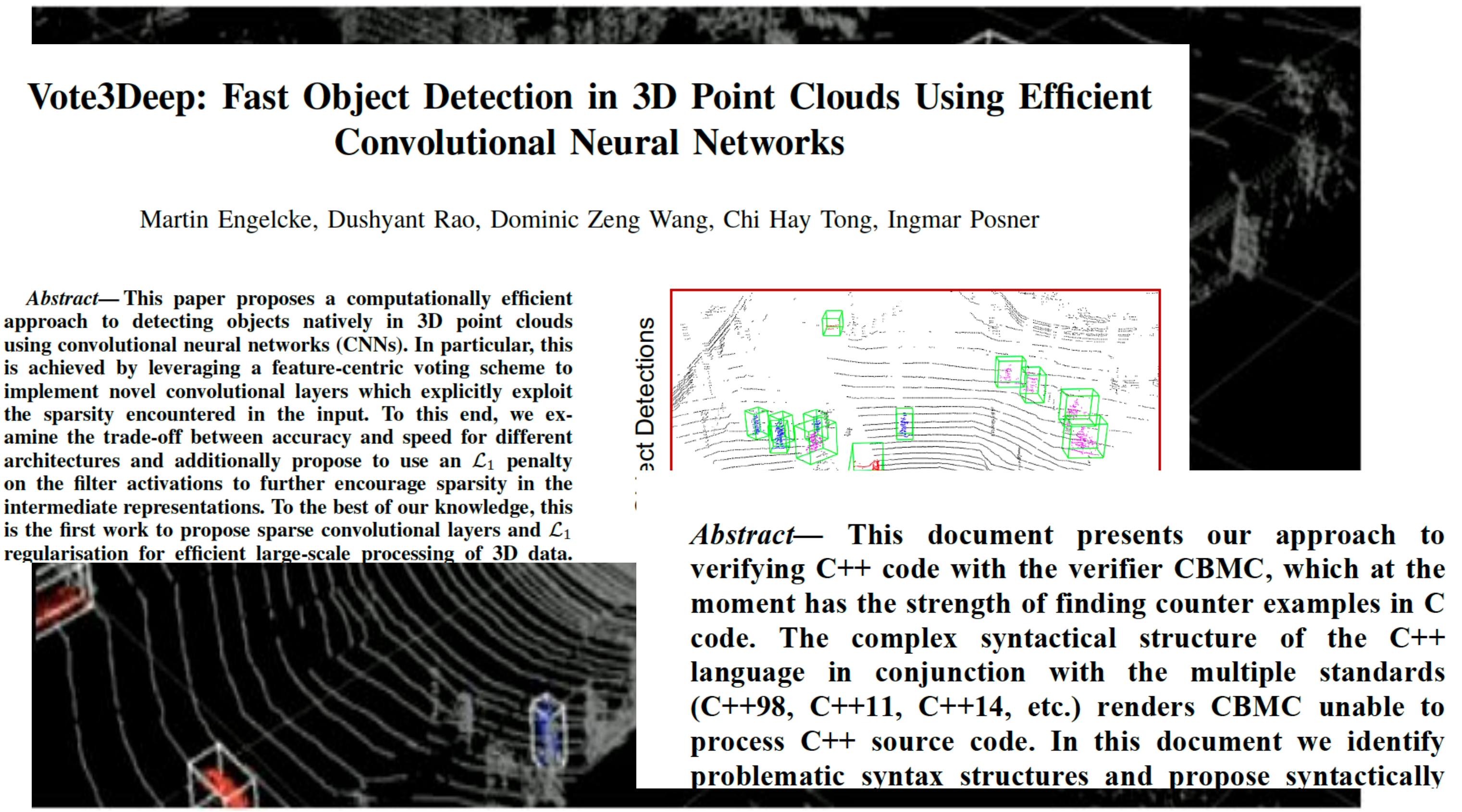
But what’s this got to do with software verification?



Self-driving in Oxford....



But what has this got to do with software verification?



New challenge: verification for ML

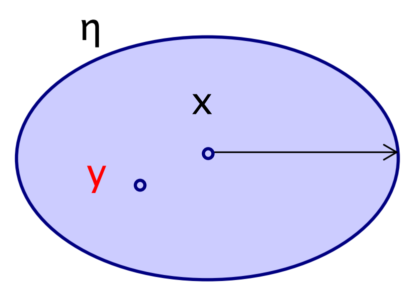
- What's different about machine learning?
  - black box, lacks interpretability
  - programming by pattern matching, not logic
  - corner cases are unseen examples, not missed conditions
  - data quality and coverage crucial
  - accuracy can be misleading
- Why is ML difficult to verify?
  - foundations of ML not well understood, mix of logic and real valued functions
  - training obscure, not clear how to choose the training method
  - dependence on choice of loss functions and optimisation
  - scalability an issue
- Need synthesis, not just verification…

This talk

- Progress in automated verification methods 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
  - journal: TCS 2019, https://arxiv.org/abs/1807.03571
- 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
  - Bayesian NN: IJCAI 2019, https://arxiv.org/abs/1903.01980

Safety of classification decisions

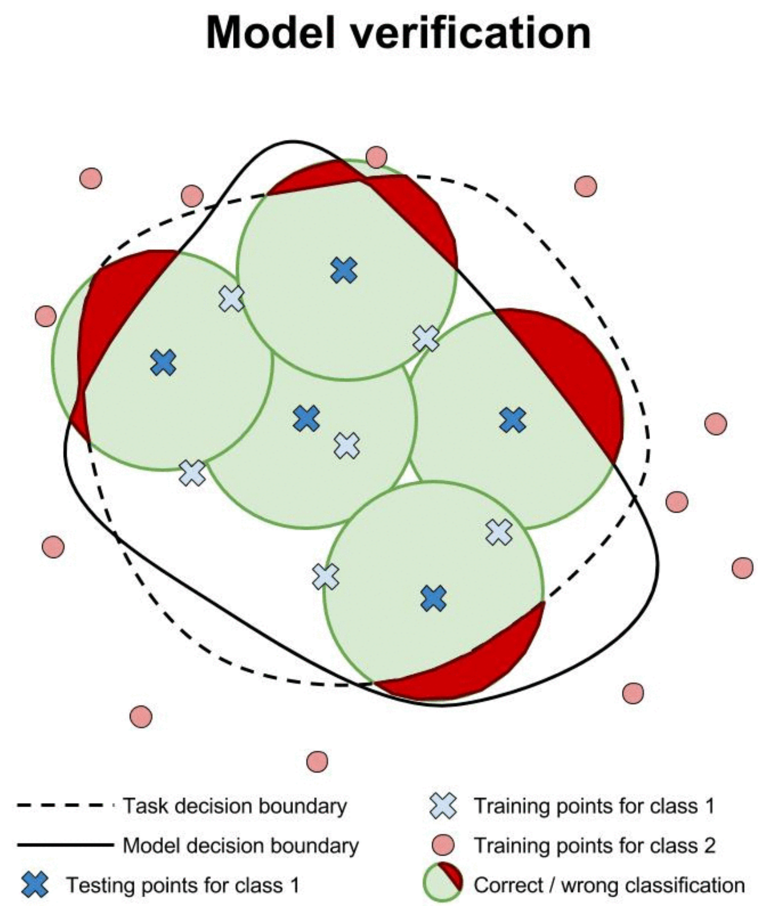
- Safety assurance process is complex
- Here focus on safety at a point as part of such a process
  - same as pointwise robustness…



- 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 over η
  - 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

Training vs testing vs verification



Safety verification

- Automated verification (= ruling out adversarial examples)
  - discretise the region, exhaustively search for misclassifications
  - provable guarantee of decision safety if adversarial 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 iterated

- But scalability remains an issues, employ various heuristics…

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

Searching for adversarial examples…

- Input space for most neural networks is high dimensional and non-linear
- Where do we start?
- How can we apply structure to the problem?

- Image of a tree has 4,000 x 2,000 x 3 dimensions = 24,000,000 dimensions
- We would like to find a very small change to these dimensions

Feature–based exploration

- Trying every combination of pixel values is intractable
- We can focus on its salient features

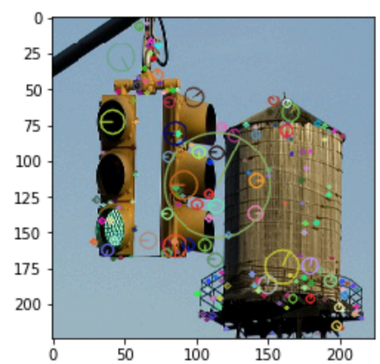
$\Lambda(\alpha)$ – Set of features given an image

λ_r – Response strength of the feature (roughly how 'important' it is)

λ_x – X coordinate of a keypoint

λ_y – Y coordinate of a keypoint

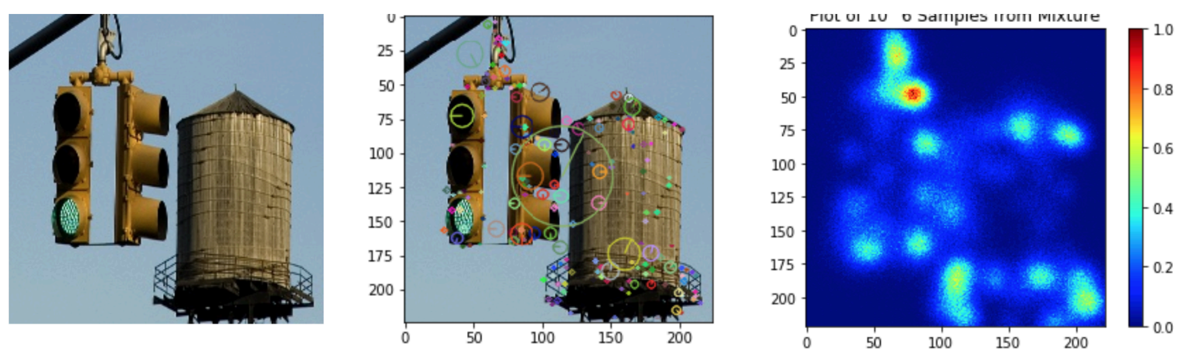
λ_s – Radius of a keypoint



Feature-based representation

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

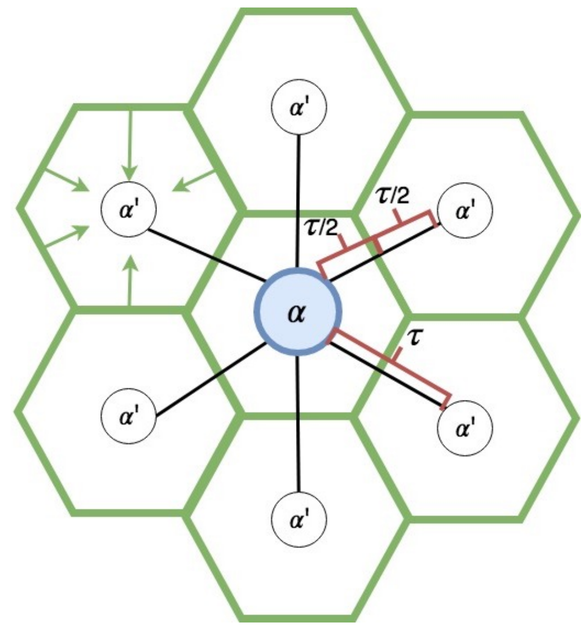
$$\mathcal{G}_{i,x} = \frac{1}{\sqrt{2\pi\lambda_{i,s}^2}} exp(\frac{-(p_x - \lambda_{i,x})^2}{2\lambda_{i,s}^2}) \quad \mathcal{G}_{i,y} = \frac{1}{\sqrt{2\pi\lambda_{i,s}^2}} exp(\frac{-(p_y - \lambda_{i,y})^2}{2\lambda_{i,s}^2})$$



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

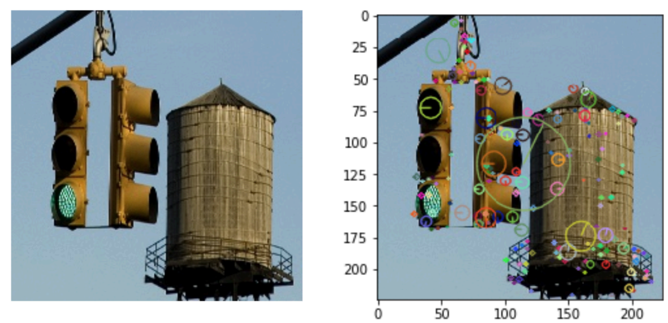
Lipschitz networks

- Lipschitz continuity limits the rate of change of output
- For Lipschitz networks, there exists a diameter such that every image within it shares the classification of a given input (smoothness)
- Use this fact to provide safety guarantees: suffices to inspect the corners of the region



Game–based search

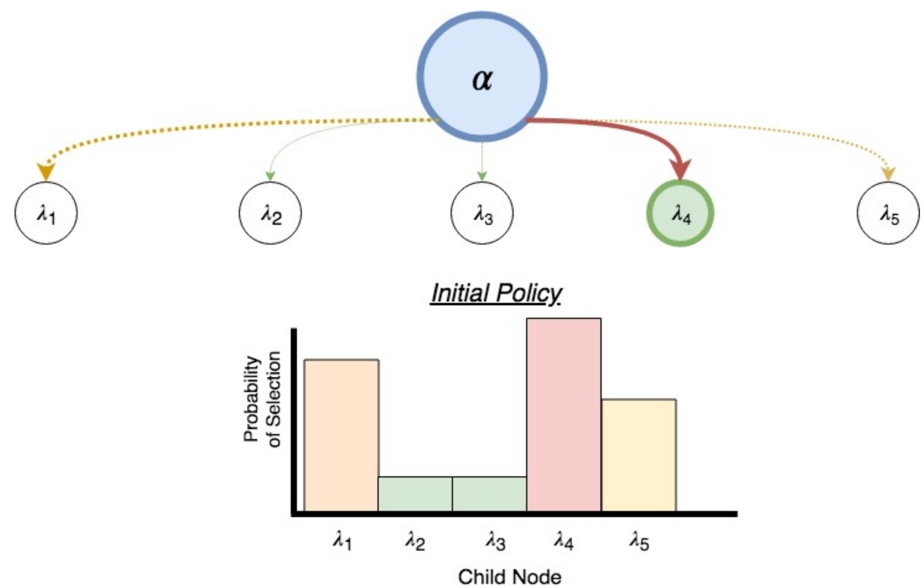
- Goal is finding adversarial example, define reward as inverse of distance
- Player 1 selects the feature that we will manipulate



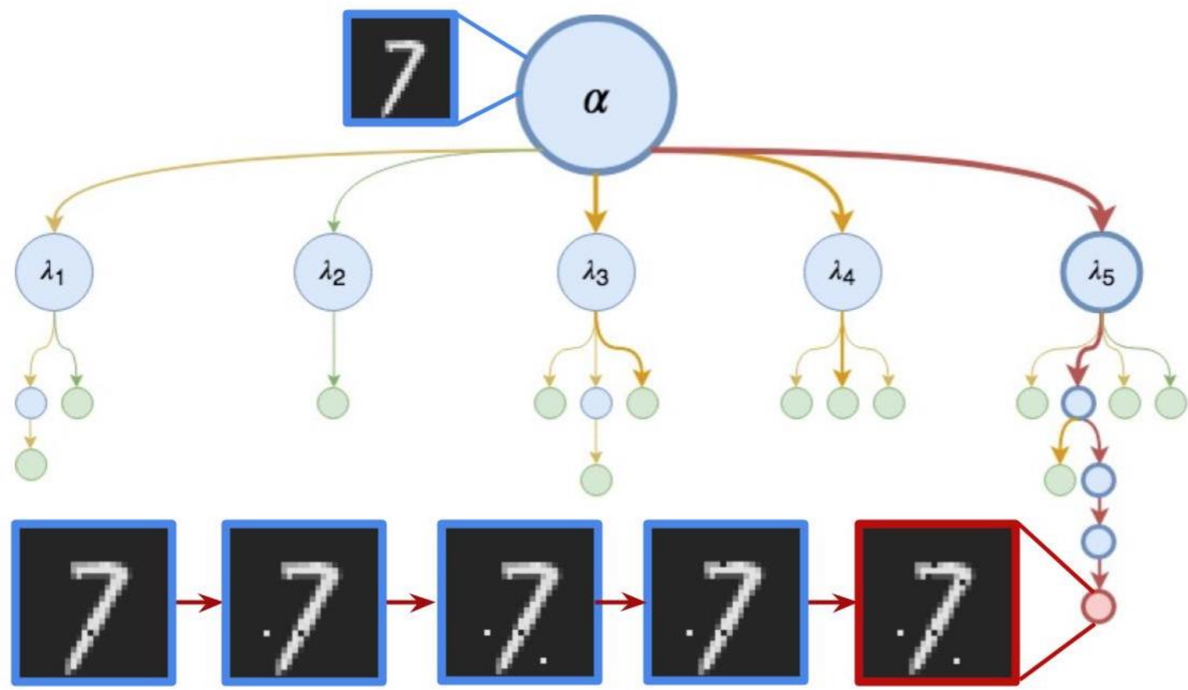
- Each feature represents a possible move for player 1
- Player 2 then selects the pixels in the feature to manipulate
- Use Monte Carlo tree search (MCTS) to explore the game tree, while querying the network to align features
- Method black/grey box, can approximate the maximum safe radius for a given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

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 previously then we expand the tree to select a new leaf

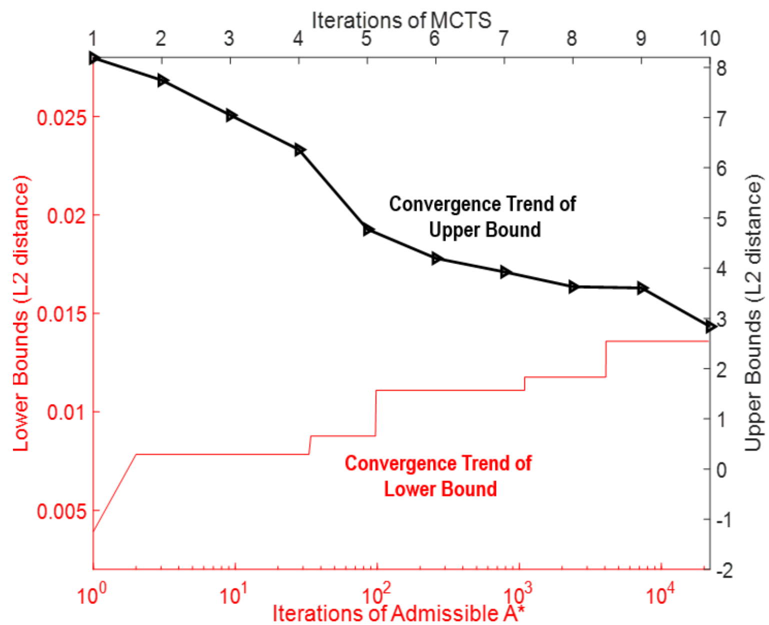


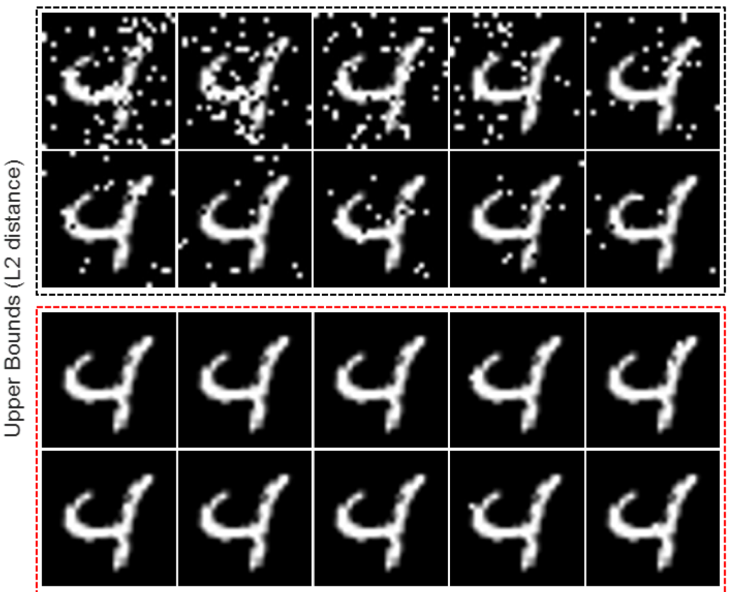
Tree expands until example is found



Now also lower bounds (MNIST)

- Convergence of lower and upper bounds on maximum safe radius





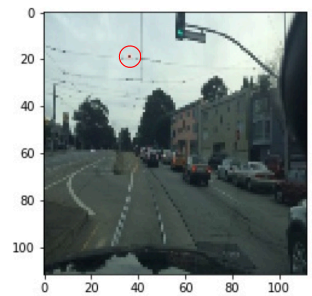
- See TCS 2019, https://arxiv.org/abs/1807.03571

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%

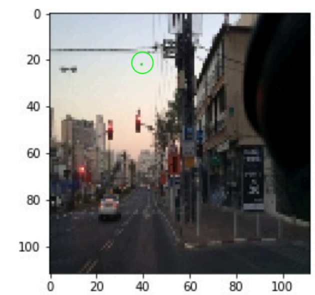
- On average, each input took less than a second to manipulate (.304 seconds)

- On average each image was vulnerable to 3 pixel changes

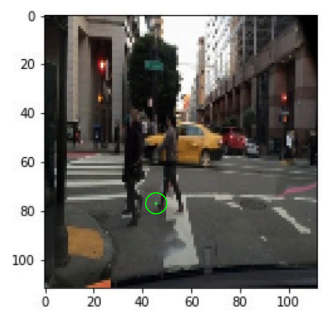


(a)

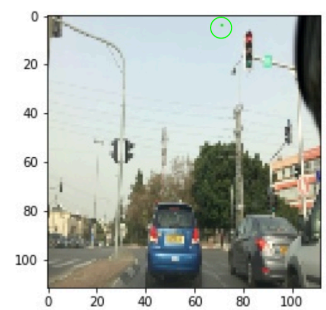
(b)



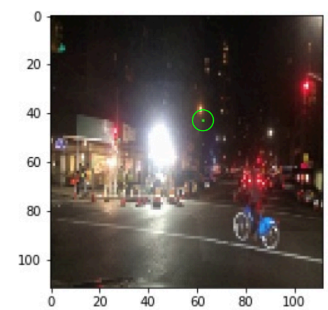
(c)



(a)



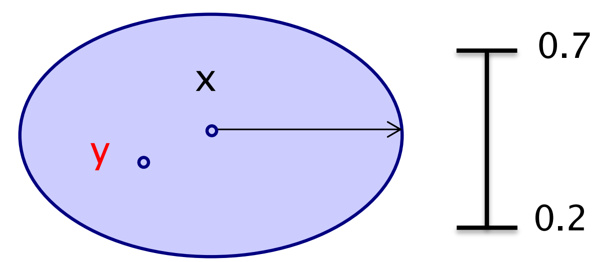
(b)



(c)

Alternative approach: reachability analysis

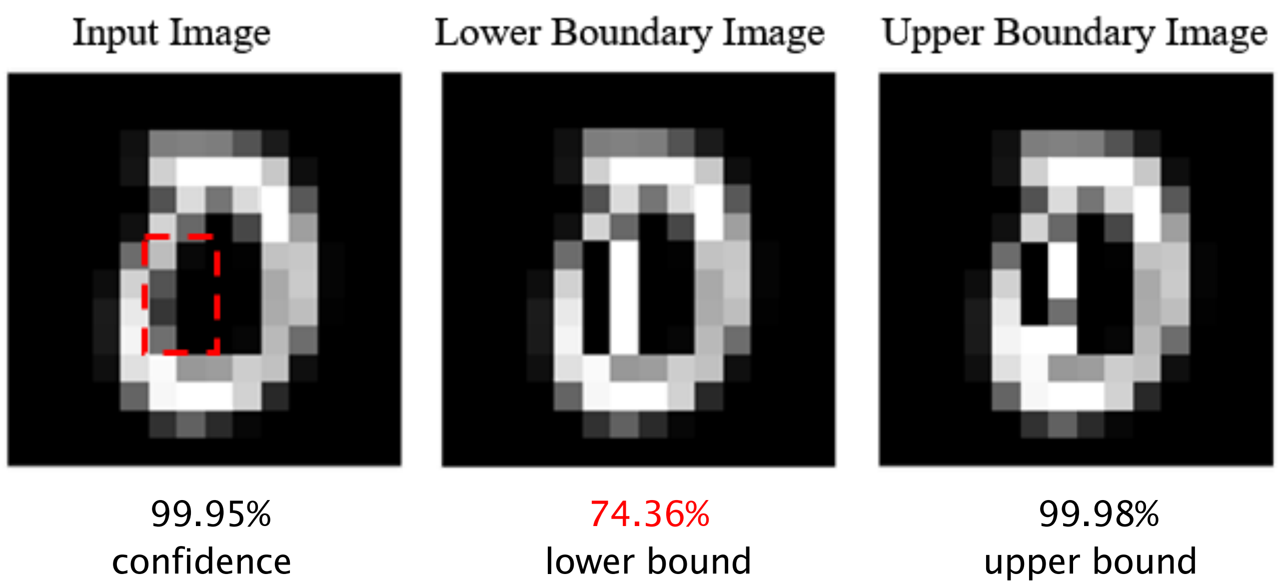
- Rather than search the discretized region, can we compute the reachable values?
- Under assumption of Lipschitz continuity
  - for $x \in \eta$, compute maximum/minimum value of $f(\eta)$
  - using global optimisation
  - anytime fashion
- Gives provable guarantees
  - best/worst case confidence values
  - 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

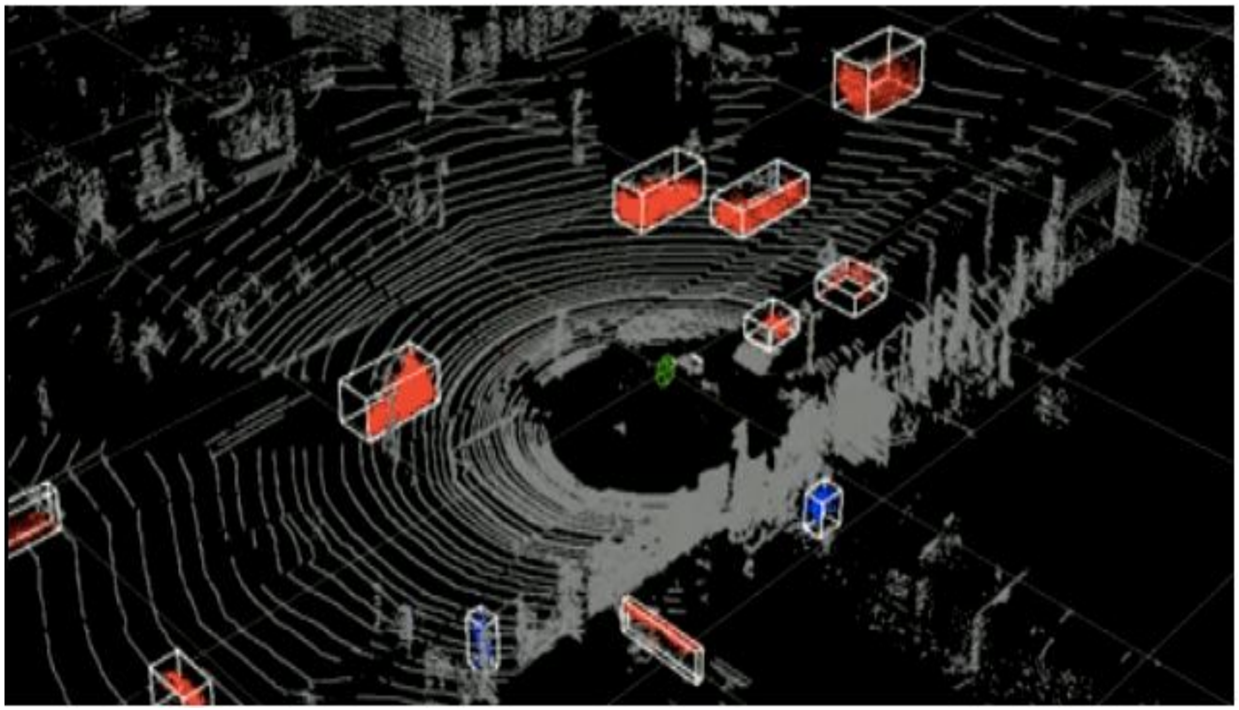
MNIST example

- Take an image and select a feature within it



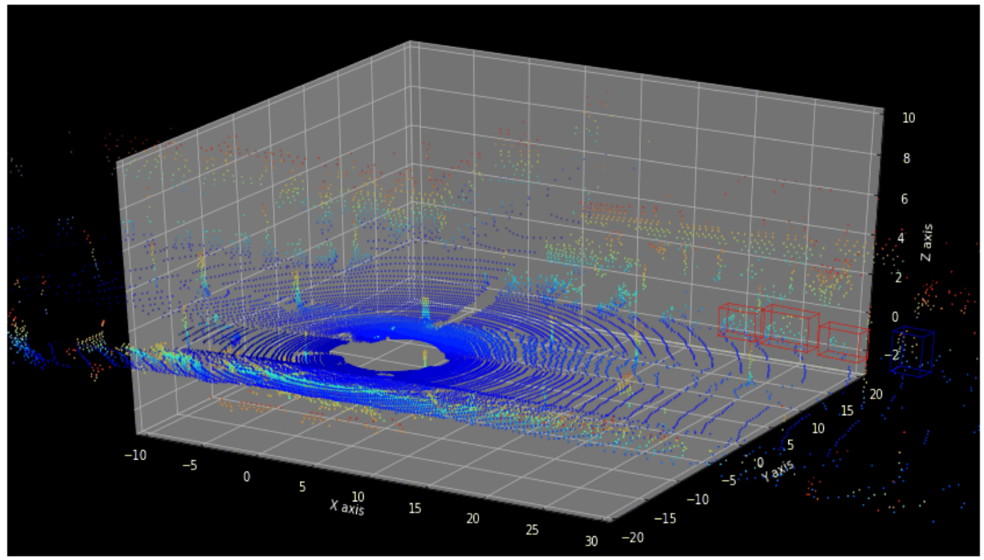
- Safety verification for the feature
  - manipulating the feature can only reduce confidence to 74.36%

Recent progress: 3D deep learning



Credits: Oxford Robotics Institute

What is LiDAR?

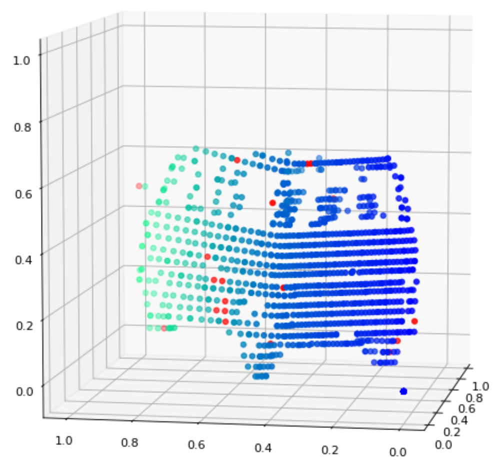


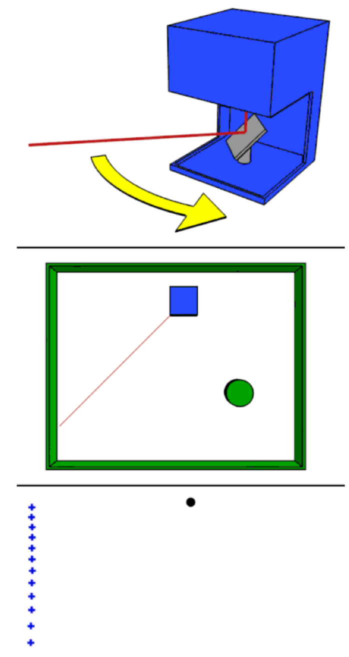
LiDAR stands for 'Light Detection And Ranging.'

Differences in laser return times and wavelengths can be used to make digital 3D representations of the environment.

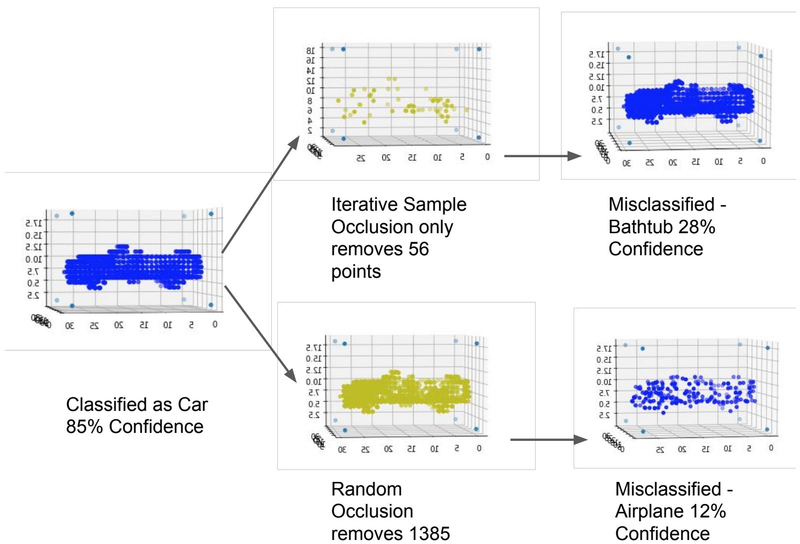
LiDAR and inherent error in point clouds

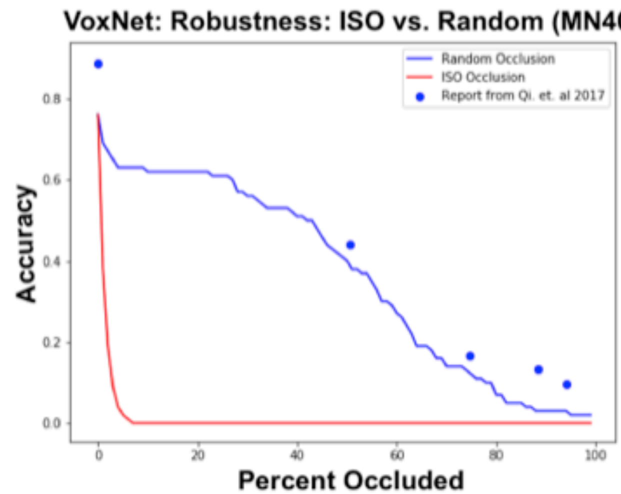
- Point ordering matters
- Partial occlusion of contiguous points
- Dark black could affect the reliability of sensor
- Misoriented sensors
- Need sub-second decision making





Can also attack 3D deep learning…





…reduce accuracy to 0% after occlusion of 6.5% of the occupied input space, targeting the critical set

Robustness of 3D Deep Learning in an Adversarial Setting. Wicker & K, In Proc. CVPR 2019.

But more progress needed…

'I hate them': Locals reportedly are frustrated with Alphabet's self-driving cars

- Alphabet's self-driving cars are said to be annoying their neighbors in Arizona, where Waymo has been testing its vehicles for the last year.
- More than a dozen locals told The Information they they hated the cars, which often struggle to cross a T-intersection near the company's office.
- The anecdotes highlight how challenging it is for self-driving cars, which are programmed to drive conservatively, to handle certain situations.

Published 3:04 PM ET Tue, 28 Aug 2018 | Updated 12:53 PM ET Wed, 29 Aug 2018

CNBC

Source: Waymo

Self-driving cars should be allowed to mount pavements and break speed limit in emergencies

share

28

A Tesla Model S

Conclusion

- Deep learning should be more critically evaluated when put into practice in safety- and security-critical situations
  - formal methods and verification have a role to play
- Overviewed methods for safety verification/testing of deep neural networks
  - search-based and feature-guided exploration, with guarantees
  - reachability computation for Lipschitz continuous networks
- Future work
  - how best to use adversarial examples: training vs logic
  - scalability
  - probabilistic guarantees
  - more complex properties
  - correct-by-construction synthesis

Acknowledgements

- My group and collaborators in this work
- Project funding
  - ERC Advanced Grant VERIWARE
  - EPSRC Mobile Autonomy Programme Grant
- See also
  - PRISM www.prismmodelchecker.org
- New ERC Advanced Grant FUN2MODEL

  "From FUNction-based TO MOdel-based automated probabilistic reasoning for DEep Learning"
- Postdoctoral and PhD positions