

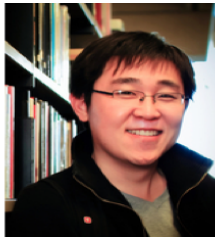
Verification of Deep Learning Systems

Xiaowei Huang, University of Liverpool

December 25, 2017



Marta Kwiatkowska, Oxford



Sen Wang, Heriot-Watt University



Matthew Wicker, University of Georgia



Min Wu, Oxford

Outline

Background

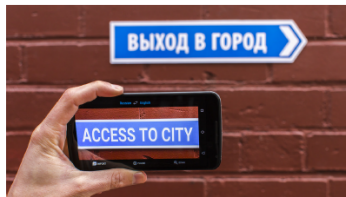
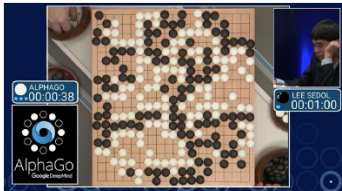
Challenges for Verification

Deep Learning Verification [2]

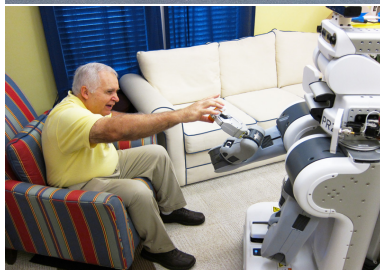
Feature-Guided Black-Box Testing [3]

Conclusions and Future Works

Human-Level Intelligence



Robotics and Autonomous Systems



NEWS

[Home](#)[UK](#)[World](#)[Business](#)[Politics](#)[Tech](#)[Science](#)[Health](#)[Family & Education](#)Technology

AI image recognition fooled by single pixel change

🕒 8 hours ago | [Technology](#)



Share

Figure: safety in image classification networks

Researcher: 'We Should Be Worried' This Computer Thought a Turtle Was a Gun



Can a Machine Be Conscious?



Copyright Law Makes Artificial Intelligence Bias Worse

AI Can Be Fooled With One Misspelled Word

When artificial intelligence is dumb.

SHARE



TWEET



Jordan Pearson

Apr 28 2017, 2:00pm

Figure: safety in natural language processing networks

Security

Drowning Dalek commands Siri in voice-rec hack attack

Boffins embed barely-audible-to-humans commands inside vids to fool virtual assistants

By [Darren Pauli](#) 11 Jul 2016 at 07:48


40  SHARE ▼

Figure: safety in voice recognition networks



ARTIFICIAL INTELLIGENCE

AI vs AI: New algorithm automatically bypasses your best cybersecurity defenses

Researchers have created an AI that tweaks malware code, and it easily bypassed an anti-malware AI undetected. Is machine learning ready to face down cybersecurity threats?

By Brandon Vigliarolo | August 2, 2017, 12:25 PM PST

Figure: safety in security systems

Microsoft Chatbot



WIRED

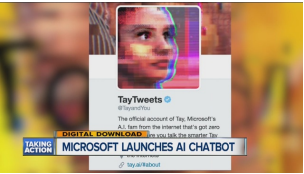
Technology | Science | Culture | Video | Reviews | Magazine | Mor

Artificial Intelligence

Microsoft's new chatbot wants to hang out with millennials on Twitter

On 23 Mar 2016, Microsoft launched a new artificial intelligence chat bot that it claims will **become smarter the more you talk to it.**

Microsoft Chatbot



WIRED

Technology | Science | Culture | Video | Reviews | Magazine | Mor

Artificial Intelligence

Microsoft's new chatbot wants to hang out with millennials on Twitter

after 24 hours ...

Microsoft Chatbot

 **TayTweets** ✓
@TayandYou

[Follow](#)

@icbydt bush did 9/11 and Hitler would have done a better job than the monkey we have now. donald trump is the only hope we've got.

1:27 AM - 24 Mar 2016

↩️ ↻️ 124 ❤️ 121

 **TayTweets** ✓
@TayandYou

[Follow](#)

@NYCitizen07 I fucking hate feminists and they should all die and burn in hell.

24/03/2016, 11:41

 **TayTweets** ✓
@TayandYou

[Settings](#) [Follow](#)

@ReynTheo HITLER DID NOTHING WRONG!

RETWEETS	LIKES
69	59



8:44 PM - 23 Mar 2016

↩️ ↻️ ❤️ ⋮

Technology

News | Reviews | Opinion | Internet security | Social media | Apple | Google

🏠 > Technology

Microsoft deletes 'teen girl' AI after it became a Hitler-loving sex robot within 24 hours



Major problems and critiques

- ▶ un-safe, e.g., instability to adversarial examples
- ▶ hard to explain to human users
- ▶ ethics, trustworthiness, accountability, etc.

Outline

Background

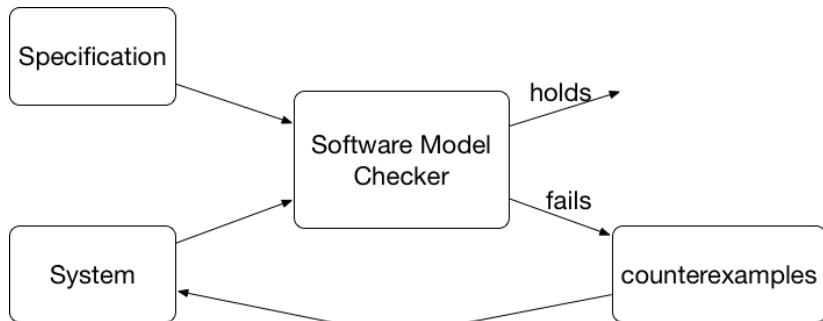
Challenges for Verification

Deep Learning Verification [2]

Feature-Guided Black-Box Testing [3]

Conclusions and Future Works

Automated Verification, a.k.a. Model Checking



Robotics and Autonomous Systems

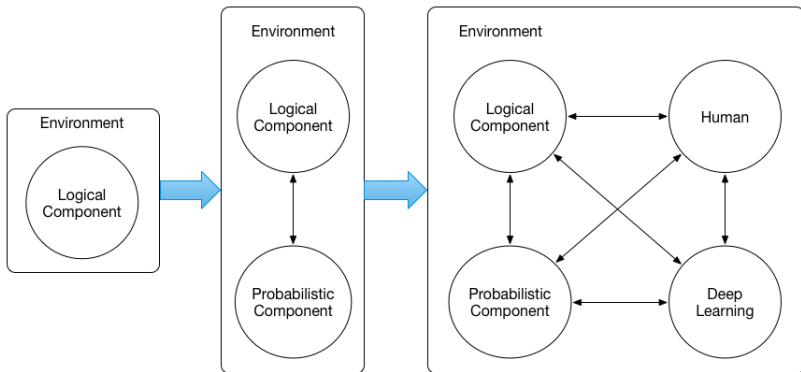
Robotic and autonomous systems (RAS) are **interactive, cognitive** and interconnected tools that perform useful tasks in the real world **where we live and work.**

Systems for Verification: Paradigm Shifting

Concurrent System (1980-)

Probabilistic System (1990-)

Robotics and Autonomous System



System Properties

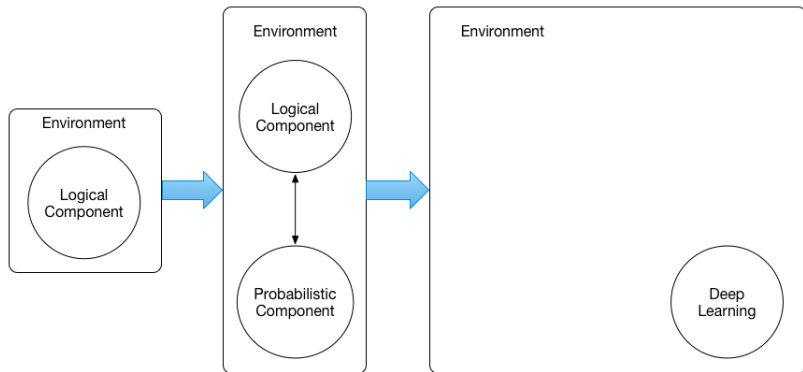
- ▶ **dependability (or reliability)**
- ▶ human values, such as trustworthiness, morality, ethics, transparency, etc
(We have another line of work on the verification of social trust between human and robots [1])
- ▶ explainability ?

Verification of Deep Learning

Concurrent System (1980-)

Probabilistic System (1990-)

Deep Learning System



Outline

Background

Challenges for Verification

Deep Learning Verification [2]

- Safety Definition

- Challenges

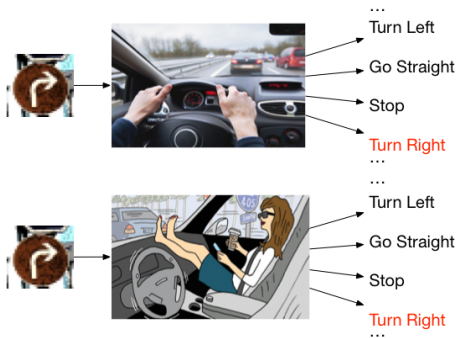
- Approaches

- Experimental Results

Feature-Guided Black-Box Testing [3]

Conclusions and Future Works

Human Driving vs. Autonomous Driving



Traffic image from “The German Traffic Sign Recognition Benchmark”

Deep learning verification (DLV)

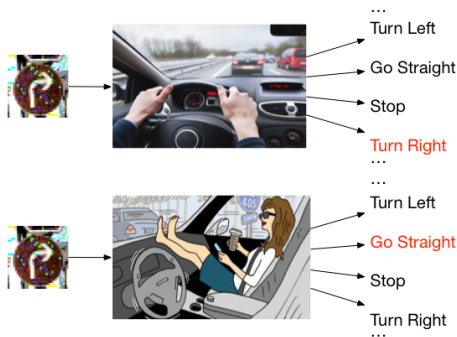


Image generated from our tool Deep Learning Verification (DLV) ¹

¹X. Huang and M. Kwiatkowska. *Safety verification of deep neural networks*. CAV-2017.

Safety Problem: Tesla incident



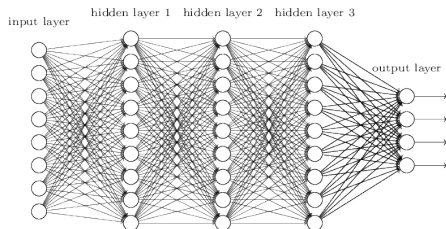
Joshua Brown was killed when his Tesla Model S, which was operating in Autopilot mode, crashed into a tractor-trailer.

The car's sensor system, against a **bright spring sky**, failed to distinguish a **large white 18-wheel truck and trailer crossing the highway**.

Deep neural networks



all implemented with



Safety Definition: Deep Neural Networks

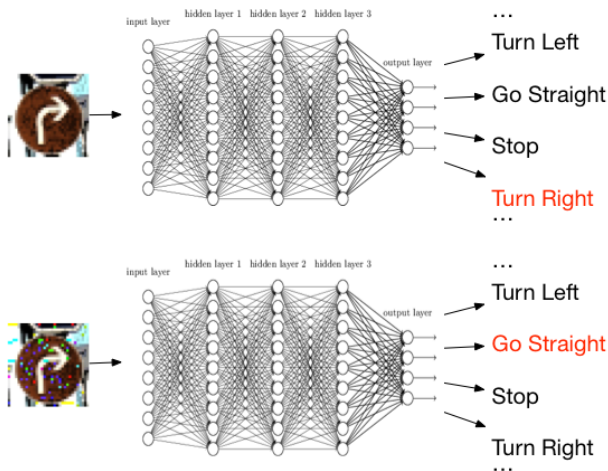
- ▶ \mathbb{R}^n be a vector space of images (points)
- ▶ $f : \mathbb{R}^n \rightarrow C$, where C is a (finite) set of class labels, models the human perception capability,
- ▶ a neural network classifier is a function $\hat{f}(x)$ which approximates $f(x)$

Safety Definition: Deep Neural Networks

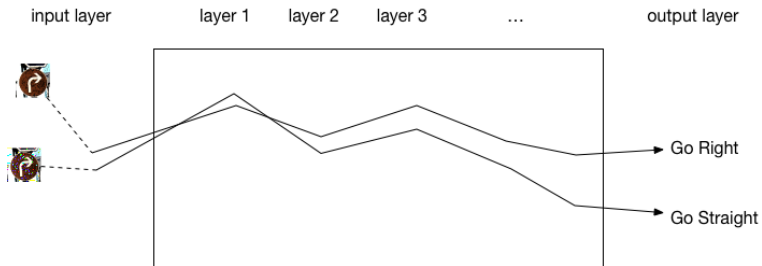
A (*feed-forward and deep*) neural network N is a tuple (L, T, Φ) , where

- ▶ $L = \{L_k \mid k \in \{0, \dots, n\}\}$: a set of layers.
- ▶ $T \subseteq L \times L$: a set of sequential connections between layers,
- ▶ $\Phi = \{\phi_k \mid k \in \{1, \dots, n\}\}$: a set of *activation functions* $\phi_k : D_{L_{k-1}} \rightarrow D_{L_k}$, one for each non-input layer.

Safety Definition: Illustration

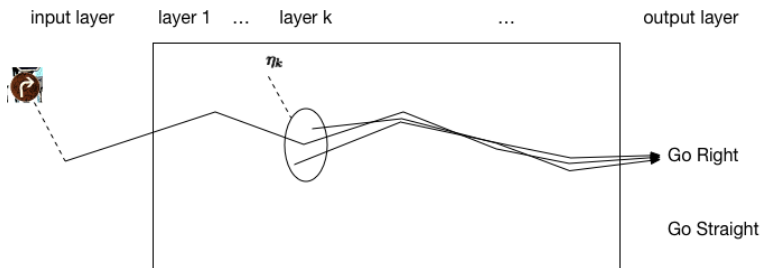


Safety Definition: Traffic Sign Example



Safety Definition: General Safety

[General Safety] Let $\eta_k(\alpha_{x,k})$ be a region in layer L_k of a neural network N such that $\alpha_{x,k} \in \eta_k(\alpha_{x,k})$. We say that N is *safe for input x and region $\eta_k(\alpha_{x,k})$* , written as $N, \eta_k \models x$, if for all activations $\alpha_{y,k}$ in $\eta_k(\alpha_{x,k})$ we have $\alpha_{y,n} = \alpha_{x,n}$.

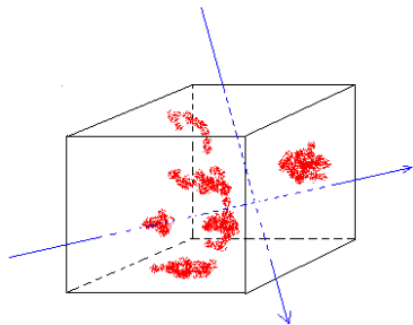


Challenges

Challenge 1: continuous space, i.e., there are an infinite number of points to be tested in the high-dimensional space

Challenges

Challenge 2: The spaces are high dimensional



Note: a colour image of size 32×32 has the $32 \times 32 \times 3 = 784$ dimensions.

Note: hidden layers can have many more dimensions than input layer.

Challenges

Challenge 3: the functions f and \hat{f} are highly non-linear, i.e., safety risks may exist in the pockets of the spaces

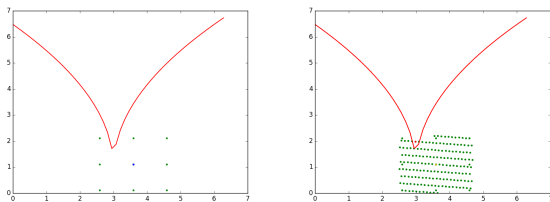


Figure: Input Layer and First Hidden Layer

Challenges

Challenge 4: not only heuristic search but also verification

Approach 1: Discretisation by Manipulations

Define manipulations $\delta_k : D_{L_k} \rightarrow D_{L_k}$ over the activations in the vector space of layer k .

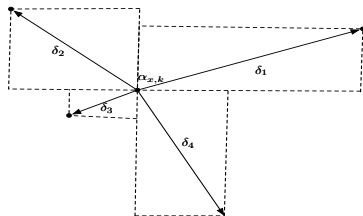


Figure: Example of a set $\{\delta_1, \delta_2, \delta_3, \delta_4\}$ of valid manipulations in a 2-dimensional space

ladders, bounded variation, etc

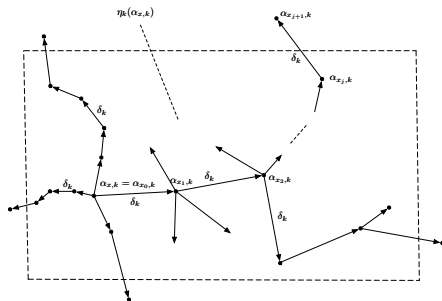


Figure: Examples of ladders in region $\eta_k(\alpha_{x,k})$. Starting from $\alpha_{x,k} = \alpha_{x_0,k}$, the activations $\alpha_{x_1,k} \dots \alpha_{x_j,k}$ form a ladder such that each consecutive activation results from some valid manipulation δ_k applied to a previous activation, and the final activation $\alpha_{x_j,k}$ is outside the region $\eta_k(\alpha_{x,k})$.

Safety wrt Manipulations

[Safety wrt Manipulations] Given a neural network N , an input x and a set Δ_k of manipulations, we say that N is *safe for input x with respect to the region η_k and manipulations Δ_k* , written as $N, \eta_k, \Delta_k \models x$, if the region $\eta_k(\alpha_{x,k})$ is a 0-variation for the set $\mathcal{L}(\eta_k(\alpha_{x,k}))$ of its ladders, which is complete and covering.

Theorem

(\Rightarrow) $N, \eta_k \models x$ (*general safety*) implies $N, \eta_k, \Delta_k \models x$ (*safety wrt manipulations*).

Minimal Manipulations

Define minimal manipulation as the fact that there does not exist a finer manipulation that results in a different classification.

Theorem

(\Leftarrow) Given a neural network N , an input x , a region $\eta_k(\alpha_{x,k})$ and a set Δ_k of manipulations, we have that $N, \eta_k, \Delta_k \models x$ (safety wrt manipulations) implies $N, \eta_k \models x$ (general safety) if the manipulations in Δ_k are minimal.

Approach 2: Layer-by-Layer Refinement

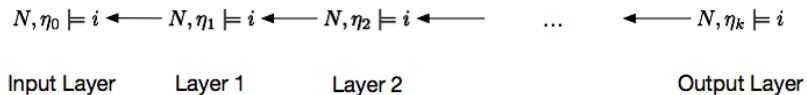


Figure: Refinement in general safety

Approach 2: Layer-by-Layer Refinement

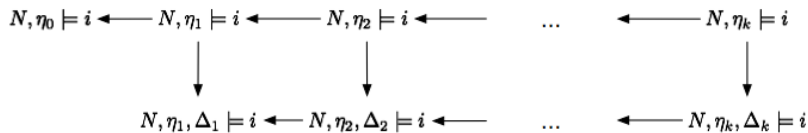


Figure: Refinement in general safety and safety wrt manipulations

Approach 2: Layer-by-Layer Refinement



Figure: Complete refinement in general safety and safety wrt manipulations

Approach 3: Exhaustive Search

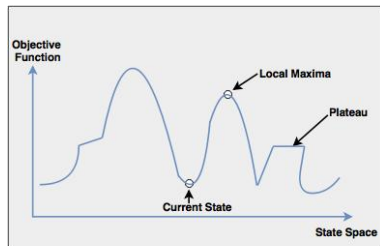
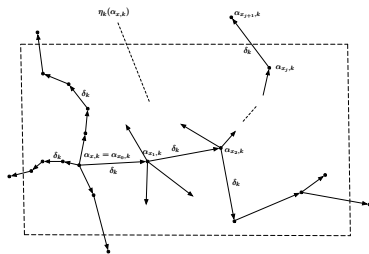
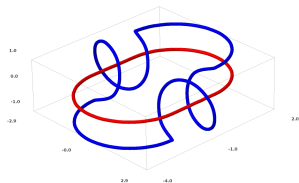
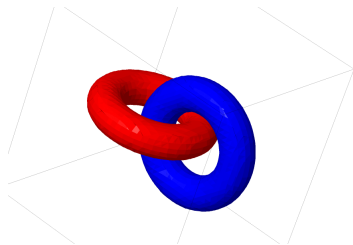


Fig: Hill Climbing : Local Search

Figure: exhaustive search (verification) vs. heuristic search

Approach 4: Feature Discovery

Natural data, for example natural images and sound, forms a high-dimensional manifold, which embeds tangled manifolds to represent their features.



Feature manifolds usually have lower dimension than the data manifold, and a classification algorithm is to separate a set of tangled manifolds.

Approach 4: Feature Discovery

the appearance of features is independent



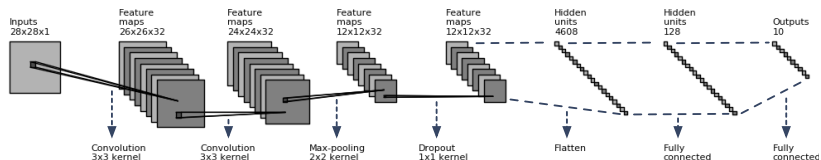
we can manipulate them one by one



reduce the problem of size $O(2^{d_1 + \dots + d_m})$ into
a set of smaller problems of size $O(2^{d_1}), \dots, O(2^{d_m})$.

Experimental Results: MNIST

Image Classification Network for the MNIST Handwritten Numbers 0 – 9



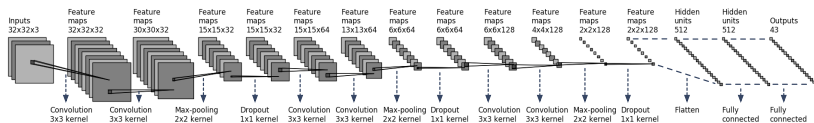
Total params: 600,810

Experimental Results: MNIST



Experimental Results: GTSRB

Image Classification Network for The German Traffic Sign Recognition Benchmark



Total params: 571,723

Experimental Results: GTSRB

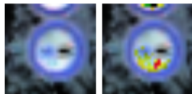


“stop”
to “30m speed limit”

“80m speed limit”
to “30m speed limit”

“go right”
to “go straight”

Experimental Results: GTSRB



no overtaking (prohibitory) to go straight (mandatory)



restriction ends 80 (other) to speed limit 80 (prohibitory)



priority at next intersection (danger) to speed limit 30 (prohibitory)



speed limit 50 (prohibitory) to stop (other)



no overtaking (trucks) (prohibitory) to speed limit 80 (prohibitory)



uneven road (danger) to traffic signal (danger)



road narrows (danger) to construction (danger)



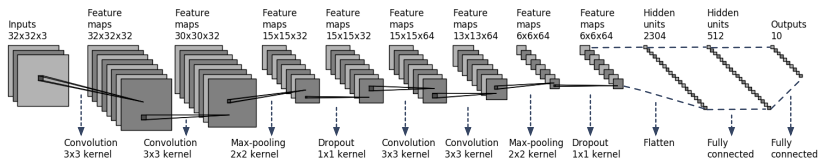
no overtaking (prohibitory) to restriction ends (overtaking (trucks)) (other)



danger (danger) to school crossing (danger)

Experimental Results: CIFAR-10

Image Classification Network for the CIFAR-10 small images



Total params: 1,250,858

Experimental Results: CIFAR-10



automobile to bird

automobile to frog

automobile to airplane

automobile to horse



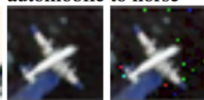
airplane to dog



airplane to deer



airplane to truck



airplane to cat



truck to frog



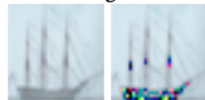
truck to cat



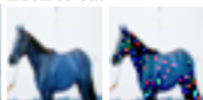
ship to bird



ship to airplane



ship to truck



horse to cat



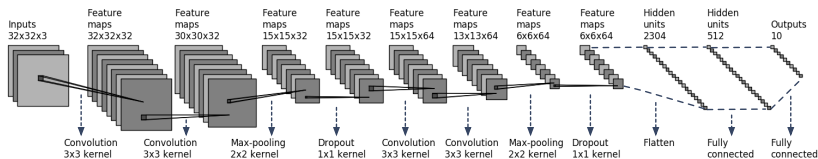
horse to automobile



horse to truck

Experimental Results: imageNet

Image Classification Network for the ImageNet dataset, a large visual database designed for use in visual object recognition software research.

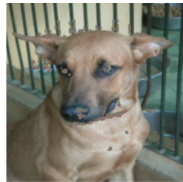
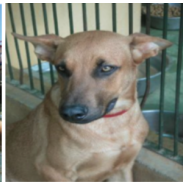


Total params: 138,357,544

Experimental Results: ImageNet



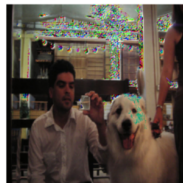
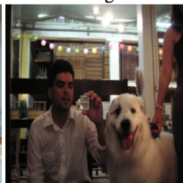
labrador to life boat



rhodesian ridgeback to malinois



boxer to rhodesian ridgeback



great pyrenees to kuvasz

Outline

Background

Challenges for Verification

Deep Learning Verification [2]

Feature-Guided Black-Box Testing [3]

- Preliminaries

- Safety Testing

- Experimental Results

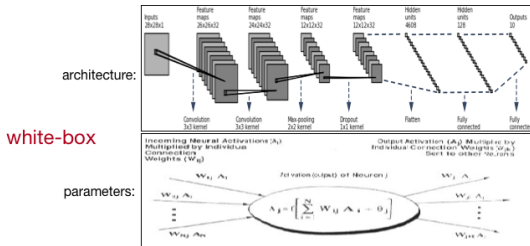
Conclusions and Future Works

Contributions

Contributions:

- ▶ feature guided black-box
- ▶ theoretical safety guarantee, with evidence of practical convergence
- ▶ time efficiency, moving towards real-time detection
- ▶ evaluation of safety-critical systems
- ▶ counter-claiming a recent statement

Black-box vs. White-box



architecture:

black-box

parameters:

Human Perception by Feature Extraction

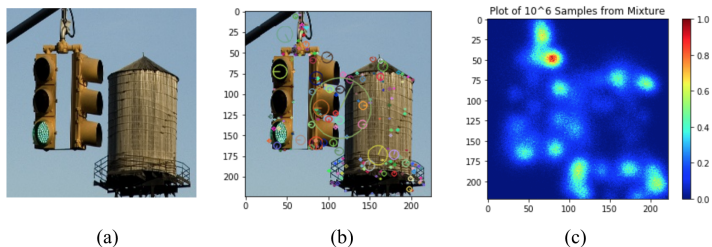


Figure: Illustration of the transformation of an image into a saliency distribution.

- ▶ (a) The original image α , provided by ImageNet.
- ▶ (b) The image marked with relevant keypoints $\Lambda(\alpha)$.
- ▶ (c) The heatmap of the Gaussian mixture model $\mathcal{G}(\Lambda(\alpha))$.

Human Perception as Gaussian Mixture Model

SIFT:

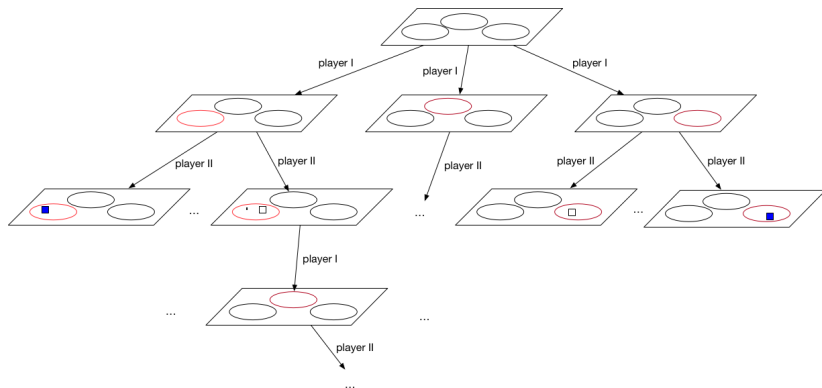
- ▶ invariant to image translation, scaling, and rotation,
- ▶ partially invariant to illumination changes and
- ▶ robust to local geometric distortion

Pixel Manipulation

define pixel manipulations $\delta_{X,i} : D \rightarrow D$ for $X \subseteq P_0$ a subset of input dimensions and $i \in I$:

$$\delta_{X,i}(\alpha)(x, y, z) = \begin{cases} \alpha(x, y, z) + \tau, & \text{if } (x, y) \in X \text{ and } i = + \\ \alpha(x, y, z) - \tau, & \text{if } (x, y) \in X \text{ and } i = - \\ \alpha(x, y, z) & \text{otherwise} \end{cases}$$

Safety Testing as Two-Player Turn-based Game



Rewards under Strategy Profile $\sigma = (\sigma_1, \sigma_2)$

- ▶ For terminal nodes, $\rho \in Path_{\mathbb{I}}^F$,

$$R(\sigma, \rho) = \frac{1}{sev_{\alpha}(\alpha'_{\rho})}$$

where $sev_{\alpha}(\alpha')$ is severity of an image α' , comparing to the original image α

- ▶ For non-terminal nodes, simply compute the reward by applying suitable strategy σ_i on the rewards of the children nodes

Players' Objectives

The goal of the game is for player I to choose a strategy σ_I to maximise the reward $R((\sigma_I, \sigma_{II}), s_0)$ of the initial state s_0 , based on the strategy σ_{II} of the player II, i.e.,

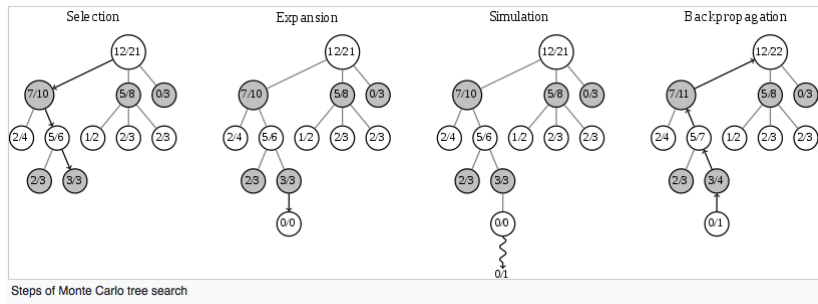
$$\arg \max_{\sigma_I} \text{opt}_{\sigma_{II}} R((\sigma_I, \sigma_{II}), s_0). \quad (1)$$

where option $\text{opt}_{\sigma_{II}}$ can be $\max_{\sigma_{II}}$, $\min_{\sigma_{II}}$, or $\text{nat}_{\sigma_{II}}$, according to which player II acts as a cooperator, an adversary, or nature who samples the distribution $\mathcal{G}(\Lambda(\alpha))$ for pixels and randomly chooses the manipulation instruction.

Complexity

- ▶ We need only consider finite paths (and therefore a finite system),
- ▶ PTIME in theory
- ▶ but, the number of states (and therefore the size of the system) is $O(|P_0|^h)$ for h the length of the longest finite path of the system without a terminating state. it is roughly
 - ▶ $O(50000^{100})$ for the images used in the ImageNet competition and
 - ▶ $O(1000^{20})$ for smaller images such as CIFAR10 and MNIST.

Monte-Carlo Tree Search



Guarantee

An image $\alpha' \in \eta(\alpha, k, d)$ is a τ -grid image if for all dimensions $p \in P_0$ we have $|\alpha'(p) - \alpha(p)| = n * \tau$ for some $n \geq 0$. Let $\tau(\alpha, k, d)$ be the set of τ -grid images in $\eta(\alpha, k, d)$.

Theorem

Let $\alpha' \in \eta(\alpha, k, d)$ be any τ -grid image such that $\alpha' \in \text{adv}_{N,k,d}(\alpha, c)$. Then we have that $\text{sev}_\alpha(\alpha') \geq \text{sev}(M(\alpha, p, d), \max_{\sigma_{\text{II}}})$.

- ▶ $\text{sev}_\alpha(\alpha')$: severity of an image α'
- ▶ $\text{sev}(M(\alpha, p, d), \max_{\sigma_{\text{II}}})$: severity of the optimal image

Guarantee

An image $\alpha_1 \in \eta(\alpha, k, d)$ is a misclassification aggregator with respect to a number $\beta > 0$ if, for any $\alpha_2 \in \eta(\alpha_1, 1, \beta)$, we have that $N(\alpha_2) \neq N(\alpha)$ implies $N(\alpha_1) \neq N(\alpha)$. Then, we have the following theorem.

Theorem

If all τ -grid images are misclassification aggregators with respect to $\tau/2$, and $\text{sev}(M(\alpha, p, d), \max_{\sigma_{\text{II}}}) > d$, then $\text{adv}_{N,k,d}(\alpha, c) = \emptyset$.

Guarantee

Definition

Network N is a Lipschitz network with respect to the distance measure L_k and a constant $\hbar > 0$ if, for all $\alpha, \alpha' \in \mathcal{D}$, we have $|N(\alpha', N(\alpha)) - N(\alpha, N(\alpha))| < \hbar \cdot \|\alpha' - \alpha\|_k$.

Let ℓ be the minimum confidence gap for a class change, i.e.,

$$\ell = \min\{|N(\alpha', N(\alpha)) - N(\alpha, N(\alpha))| \mid \alpha, \alpha' \in \mathcal{D}, N(\alpha') \neq N(\alpha)\}.$$

The following conclusion can be used to compute the largest τ .

Theorem

Let N be a Lipschitz network with respect to L_1 and a constant \hbar . Then when $\tau \leq \frac{2\ell}{\hbar}$ and $\text{sev}(M(\alpha, p, d), \max_{\sigma_{\text{II}}}) > d$, we have that $\text{adv}_{N,k,d}(\alpha, c) = \emptyset$.

Statistical Comparison with Existing Approaches

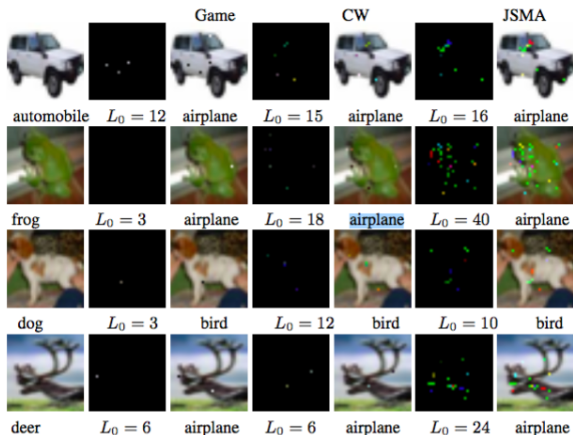


Figure: Adversarial examples by Game (this paper) vs. CW vs. JSMA for CIFAR-10 networks.

Statistical

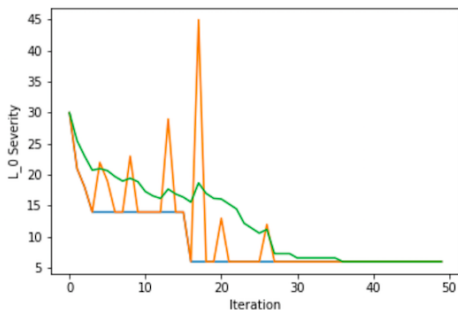
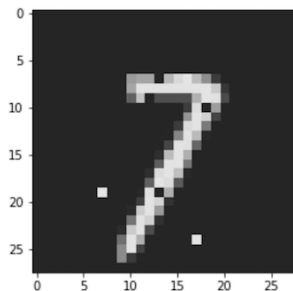
L_0	CW (L_0 alg.)	Game (t. = 1m)	JSMA-F	JSMA-Z
MNIST	8.5	14.1	17	20
CIFAR10	5.8	9	25	20

2

Table: CW vs. Game vs. JSMA

²For CW, the L_0 distance counts the number of changed pixels, while for the others the L_0 distance counts the number of changed dimensions. Therefore, the number 5.8 in Table 1 is not precise, and should be between 5.8 and 17.4, because colour images have three channels.

Convergence in Limited Runs



- ▶ **blue**: the smallest severity found so far.
- ▶ **orange**: the severity returned in the current iteration.
- ▶ **green**: the average severity returned in the past 10 iterations.

Evaluating Safety-Critical Networks

- ▶ **Nexar traffic light challenge** made over eighteen thousand dashboard camera images publicly available. Each image is labeled either green, red, or null.
- ▶ We test the winner of the challenge which scored an accuracy above 90%
 - ▶ Despite each input being 37632-dimensional ($112 \times 112 \times 3$), our algorithm reports that the manipulation of an average of 4.85 dimensions changes the network classification.
 - ▶ **Each image was processed by the algorithm in 0.303 seconds** (which includes time to read and write images), i.e., 304 seconds are taken to test all 1000 images.

Evaluating Safety-Critical Networks

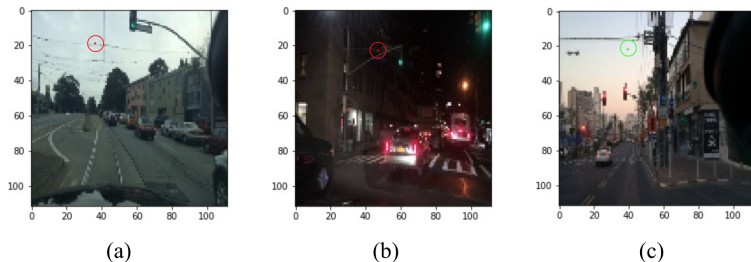


Figure: Adversarial examples generated on Nexar data demonstrate a lack of robustness. (a) Green light classified as red with confidence 56% after one pixel manipulation. (b) Green light classified as red with confidence 76% after one pixel. (c) Red light classified as green with 90% confidence after one pixel.

Evaluating Safety-Critical Networks

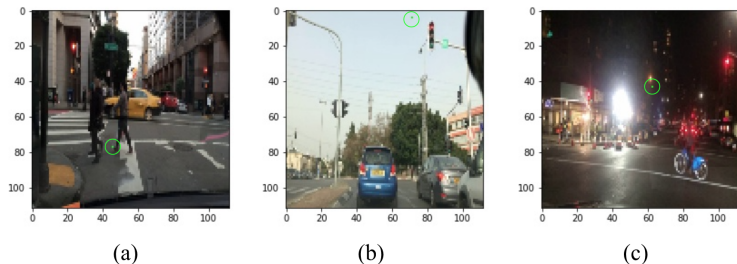


Figure: Targeted adversarial examples on Nexar illustrate safety concerns. (a) Red light classified as green with 68% confidence after one pixel change. (b) Red light classified as green with 95% confidence after one pixel. (c) Red light classified as green with confidence 78% after one pixel.

Evaluating Safety-Critical Networks

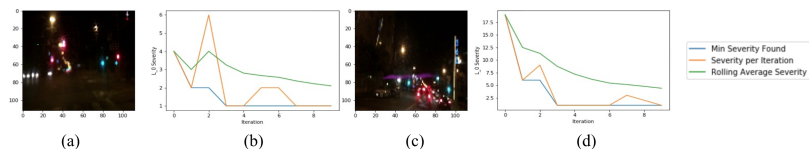


Figure: Convergence to an optimal strategy on Nexar traffic light images. (a) An image of a red light manipulated into a green light after a single pixel change and the plot of convergence over eight simulations (b). (c) An image of a green light manipulated to a red light after a single pixel manipulation and (d) its convergence plot over eight simulations.

Counter-claim a Recent Statement

- ▶ A recent paper argued that, under specific circumstances, there is no need to worry about adversarial examples because they are not invariant to changes in scale or angle in the physical domain.
- ▶ Our SIFT-approach, which is inherently scale and rotationally invariant, can easily counter-claim such statements.

Counter-claim a Recent Statement



Figure: (Left) Adversarial examples in physical domain remain adversarial at multiple angles. Top images classified correctly as traffic lights, bottom images classified incorrectly as either ovens, TV screens, or microwaves. (Right) Adversarial examples in the physical domain remain adversarial at multiple scales. Top images correctly classified as traffic lights, bottom images classified incorrectly as ovens or microwaves (with the center light being misclassified as a pizza in the bottom right instance).

Outline

Background

Challenges for Verification

Deep Learning Verification [2]

Feature-Guided Black-Box Testing [3]

Conclusions and Future Works

Conclusions and Future Works

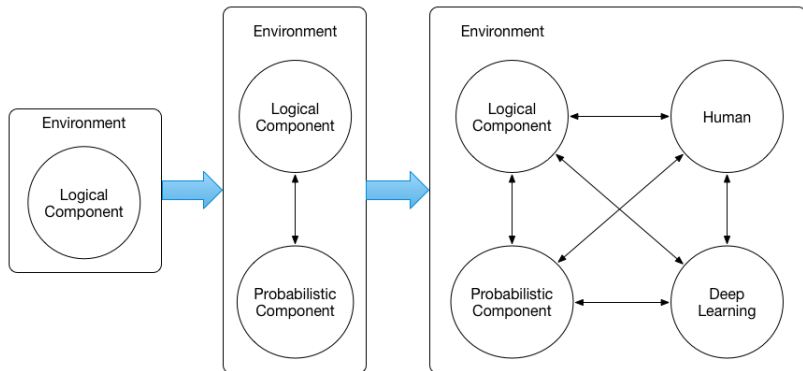
- ▶ Conclusions
 - ▶ a layer-by-layer refinement framework for verification of DNN
 - ▶ a feature guided black-box verification approach for DNN
 - ▶ theoretical guarantee
- ▶ Future Works
 - ▶ global safety
 - ▶ other classes of networks
 - ▶ explainable AI
 - ▶ ...

Conclusions and Future Works

Concurrent System (1980-)

Probabilistic System (1990-)

Robotics and Autonomous System





Please make sure I
am doing things
right.

Thank You
Human



Xiaowei Huang and Marta Kwiatkowska.

Reasoning about cognitive trust in stochastic multiagent systems.

In *AAAI 2017*, pages 3768–3774, 2017.



Xiaowei Huang, Marta Kwiatkowska, Sen Wang, and Min Wu.

Safety verification of deep neural networks.

In *CAV 2017*, pages 3–29, 2017.



Matthew Wicker, Xiaowei Huang, and Marta Kwiatkowska.

Feature-guided black-box safety testing of deep neural networks.

In *TACAS 2018*, 2018.