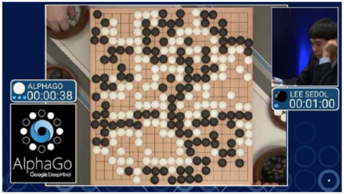
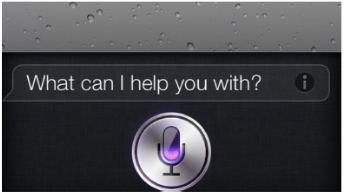
Approximate Verification of Deep Neural Networks with Provable Guarantees

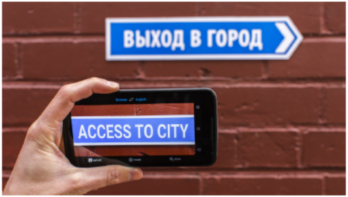
Xiaowei Huang, University of Liverpool

Outline

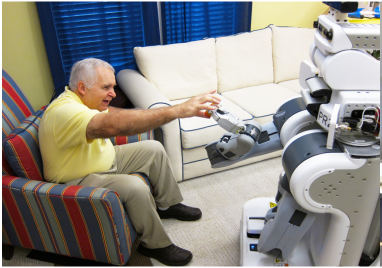
Human-Level Intelligence



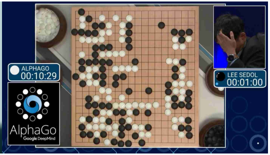




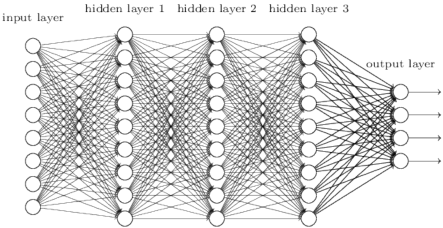
Robotics and Autonomous Systems



Deep neural networks



all implemented with



Major problems and critiques

- un-safe, e.g., lack of robustness (this talk)
- hard to explain to human users
- ethics, trustworthiness, accountability, etc.

Figure: safety in image classification networks

MOTHERBOARD

TIFICIAL INTELLIGENCE

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


The Register®
Biting the hand that feeds IT

RE SOFTWARE SECURITY TRANSFORMATION DEVOPS BUSINESS PERSONAL TECH

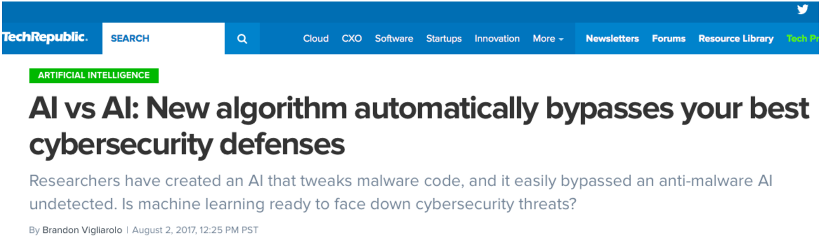
Security

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

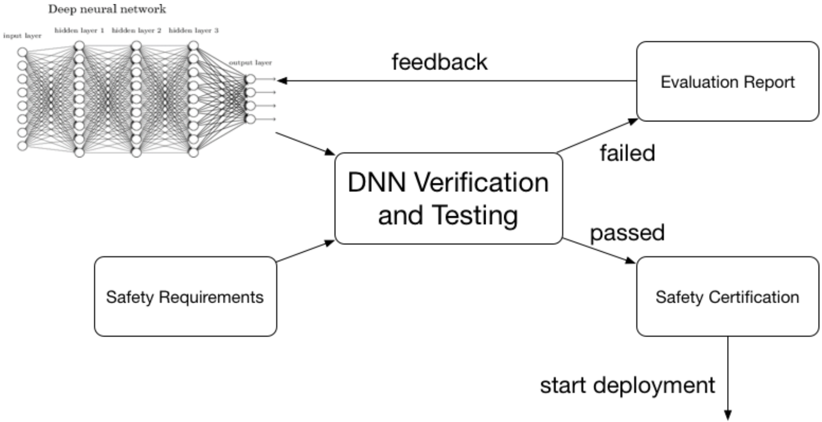
Researchers have created an AI that tweaks malware code, and it easily bypassed an anti-malware AI undetected. Is machine learning ready to face down cybersecurity threats?

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

Figure: safety in security systems

Outline

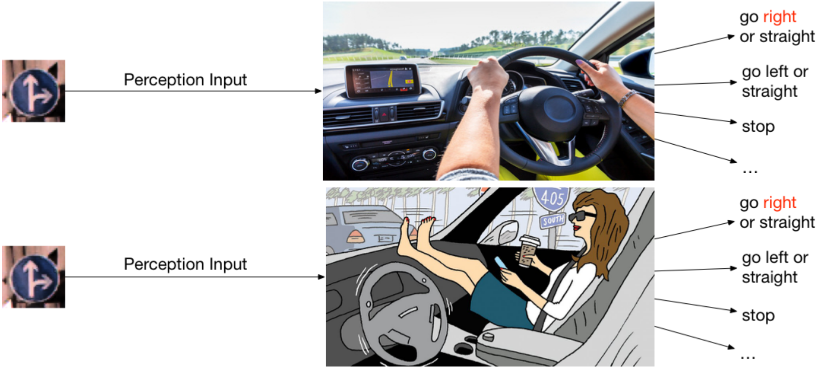
Certification of DNN



Safety Requirements

- Pointwise Robustness (this talk)
  - if the decision of a pair (input, network) is invariant with respect to the perturbation to the input.
- Network Robustness
- or more fundamentally, Lipschitz continuity, mutual information, etc
- model interpretability

Safety Definition: Human Driving vs. Autonomous Driving



Traffic image from "The German Traffic Sign Recognition Benchmark"

Safety Definition: Human Driving vs. Autonomous Driving

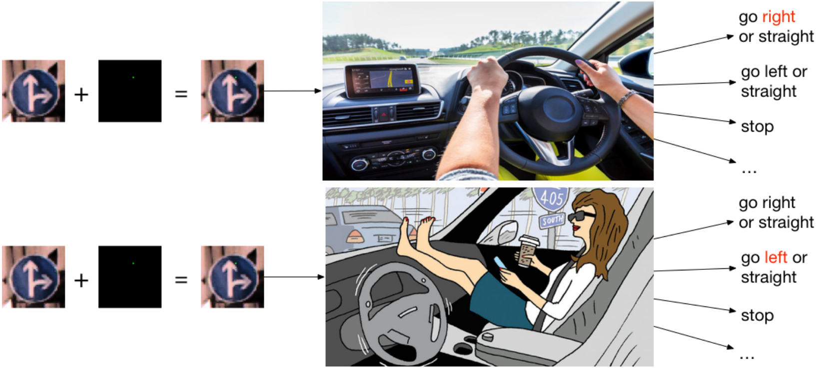
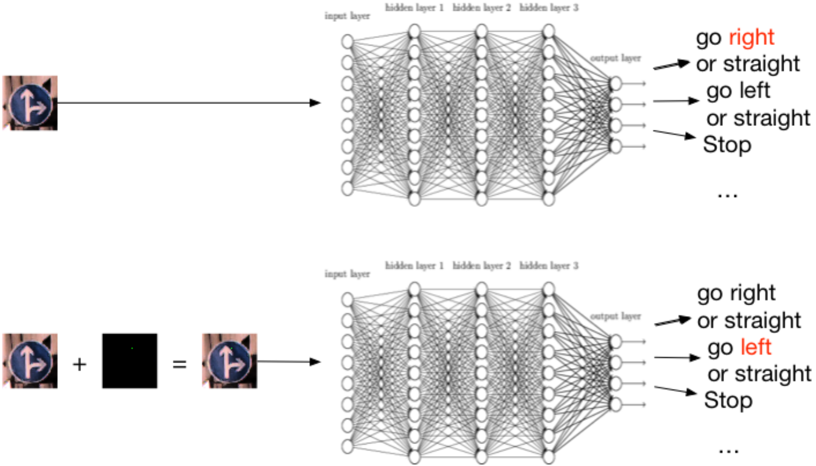


Image generated from our tool

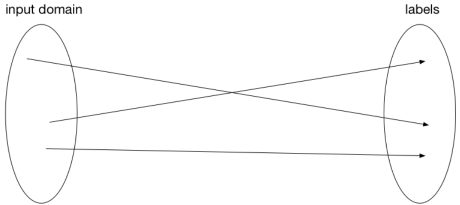
Safety Problem: Incidents

Safety Definition: Illustration



Safety Definition: Deep Neural Networks

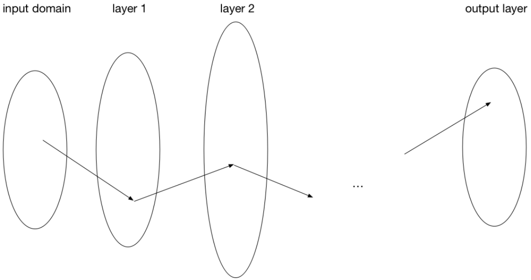
- $\mathbb{R}^n$ be a vector space of inputs (points)
- $f : \mathbb{R}^n \to C$, where C is a (finite) set of class labels, models the human perception capability,
- a neural network classifier is a function $\hat{f}(x)$ which approximates $f(x)$



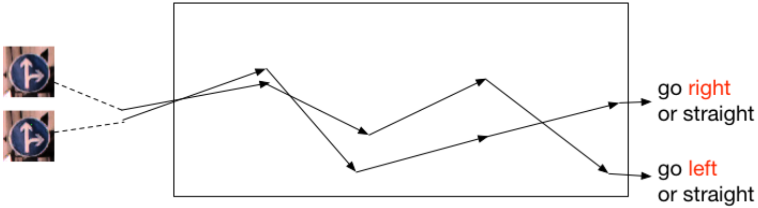
Safety Definition: Deep Neural Networks

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

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



Safety Definition: Traffic Sign Example

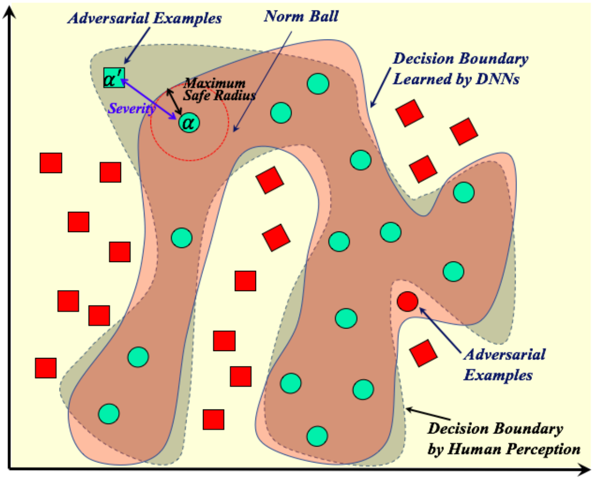


Maximum Safe Radius

Definition

The *maximum safe radius* problem is to compute the minimum distance from the original input α to an adversarial example, i.e.,

$$\mathtt{MSR}(\alpha) = \min_{\alpha' \in \mathrm{D}} \{||\alpha - \alpha'||_k \mid \alpha' \text{ is an adversarial example}\} \quad (1)$$


Adversarial Examples
Norm Ball
Decision Boundary
Learned by DNNs
α'
Maximum
Safe Radius
Severity
α
Adversarial
Examples
Decision Boundary
by Human Perception

Challenges

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

Challenge 2: The spaces are high dimensional

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

Challenge 4: not only heuristic search but also verification

Approach 1: Single Layer – Discretisation

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

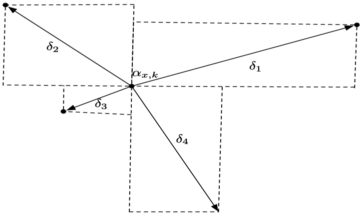
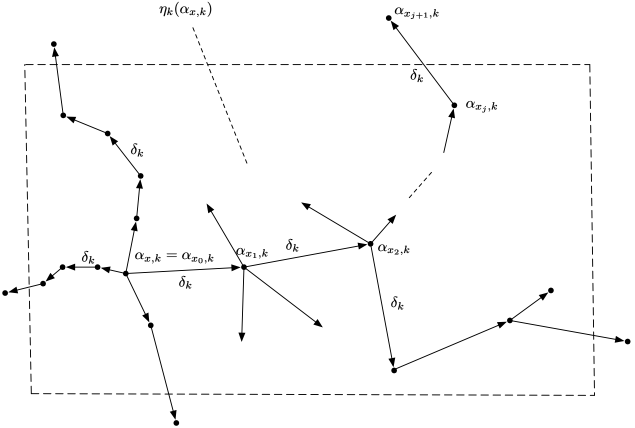


Figure: Example of a set $\{\delta_1, \delta_2, \delta_3, \delta_4\}$ of valid manipulations in a 2-dimensional space

Exploring a Finite Number of Points



Finite Approximation

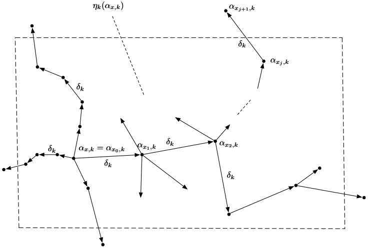
Definition

Let $\tau \in (0, 1]$ be a manipulation magnitude. The *finite maximum safe radius* problem $\texttt{FMSR}(\tau, \alpha)$ is defined over the manipulation magnitude τ (details to be given later).

Lemma

For any $\tau \in (0, 1]$*, we have that* $\texttt{MSR}(\alpha) \leq \texttt{FMSR}(\tau, \alpha)$.

Approach 2: Single Layer – Exhaustive Search



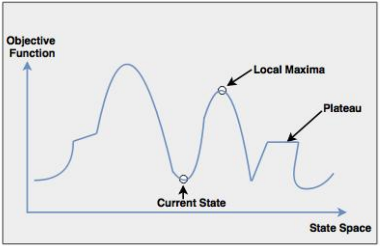
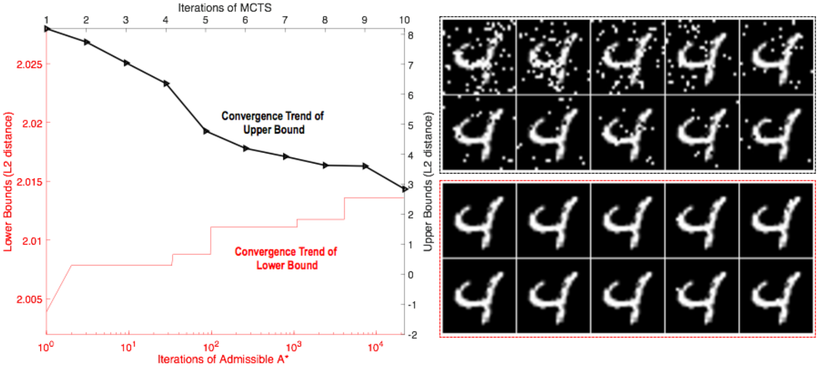


Fig: Hill Climbing : Local Search

Figure: exhaustive search (verification) vs. heuristic search

Approach 3: Single Layer – Anytime Algorithms



Approach 4: Layer-by-Layer Refinement

Input Layer | Layer 1 | Layer 2 | Output Layer

MSR_0

$\leq$

$FMSR_0(\tau_0)$

where $\tau_0 > \tau_0*$

Will explain how to determine τ_0^* later.

Approach 2: Layer-by-Layer Refinement

Input Layer | Layer 1 | Layer 2 | Output Layer

MSR_0

$\leq$

$FMSR_0(\tau_0) \geq FMSR_1(\tau_1)$

where $\tau_1 > \tau_1*$

Approach 2: Layer-by-Layer Refinement

Input Layer		Layer 1		Layer 2	...	Layer k	Output Layer
MSR_0				$=$		MSR_k	
$\leq$						$=$	
$FMSR_0(\tau_0)$	$\geq$	$FMSR_1(\tau_1)$			$\geq$	$FMSR_k(\tau_k)$	
						where $\tau_k \leq \tau_k*$	

Outline

Preliminaries: Lipschitz network

Definition

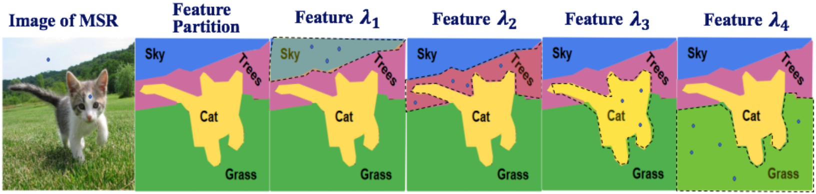
Network N is a Lipschitz network with respect to distance function L_k if there exists a constant $\hbar_c > 0$ for every class $c \in C$ such that, for all $\alpha, \alpha' \in \mathrm{D}$, we have

$$|N(\alpha', c) - N(\alpha, c)| \leq \hbar_c \cdot ||\alpha' - \alpha||_k. \tag{2}$$

Most known types of layers, including fully-connected, convolutional, ReLU, maxpooling, sigmoid, softmax, etc., are Lipschitz continuous [4].

Preliminaries: Feature-Based Partitioning

Partition the input dimensions with respect to a set of features. Here, features in the simplest case can be a uniform partition, i.e., do not necessarily follow a particular method.



Useful for the reduction to two-player game, in which player One chooses a feature and player Two chooses how to manipulate the selected feature.

Preliminaries: Input Manipulation

Let $\tau > 0$ be a positive real number representing the manipulation magnitude, then we can define *input manipulation* operations $\delta_{\tau,X,i} : \mathrm{D} \to \mathrm{D}$ for $X \subseteq P_0$, a subset of input dimensions, and $i : P_0 \to \mathbb{N}$, an instruction function by:

$$\delta_{\tau,X,i}(\alpha)(j) = \begin{cases} \alpha(j) + i(j) * \tau, & \text{if } j \in X \\ \alpha(j), & \text{otherwise} \end{cases}$$

for all $j \in P_0$.

Approximation Based on Finite Optimisation

Definition

Let $\tau \in (0,1]$ be a manipulation magnitude. The *finite maximum safe radius* problem $\texttt{FMSR}(\tau, \alpha)$ based on input manipulation is as follows:

$$\min_{\Lambda' \subseteq \Lambda(\alpha)} \min_{X \subseteq \bigcup_{\lambda \in \Lambda'} P_\lambda} \min_{i \in \mathcal{I}} \{ ||\alpha - \delta_{\tau,X,i}(\alpha)||_k \mid \delta_{\tau,X,i}(\alpha) \text{ is an adv. example} \} \tag{3}$$

Lemma

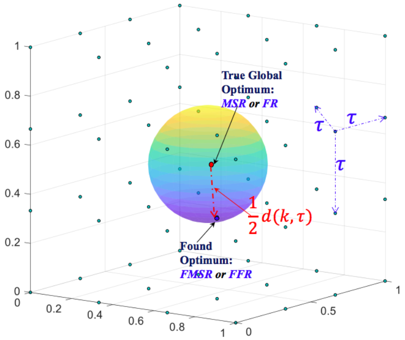
For any $\tau \in (0,1]$*, we have that* $\texttt{MSR}(\alpha) \leq \texttt{FMSR}(\tau, \alpha)$*.*

We need to determine the condition for τ to satisfy so that $\texttt{FMSR}(\tau, \alpha) = \texttt{MSR}(\alpha)$.

Grid Space

Definition

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



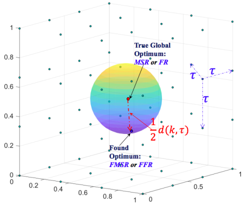
misclassification aggregator

Definition

An input $\alpha_1 \in \eta(\alpha, L_k, d)$ is a *misclassification aggregator* with respect to a number $\beta > 0$ if, for any $\alpha_2 \in \eta(\alpha_1, L_k, \beta)$, we have that $N(\alpha_2) \neq N(\alpha)$ implies $N(\alpha_1) \neq N(\alpha)$.

Lemma

If all τ-grid inputs are misclassification aggregators with respect to $\frac{1}{2}d(k,\tau)$, then $\mathtt{MSR}(k,d,\alpha,c) \geq \mathtt{FMSR}(\tau,k,d,\alpha,c) - \frac{1}{2}d(k,\tau)$.



Conditions for Achieving Misclassification Aggregator

Given a class label c, we let

$$g(\alpha', c) = \min_{c' \in C, c' \neq c} \{N(\alpha', c) - N(\alpha', c')\} \tag{4}$$

be a function maintaining for an input α' the *minimum confidence margin* between the class c and another class $c' \neq N(\alpha')$.

Lemma

Let N be a Lipschitz network with a Lipschitz constant $\hbar_c$ for every class $c \in C$. If

$$d(k, \tau) \leq \frac{2g(\alpha', N(\alpha'))}{\max_{c \in C, c \neq N(\alpha')}(\hbar_{N(\alpha')} + \hbar_c)} \tag{5}$$

for all τ-grid input $\alpha' \in G(\alpha, k, d)$, then all τ-grid inputs are misclassification aggregators with respect to $\frac{1}{2}d(k, \tau)$.

Main Theorem

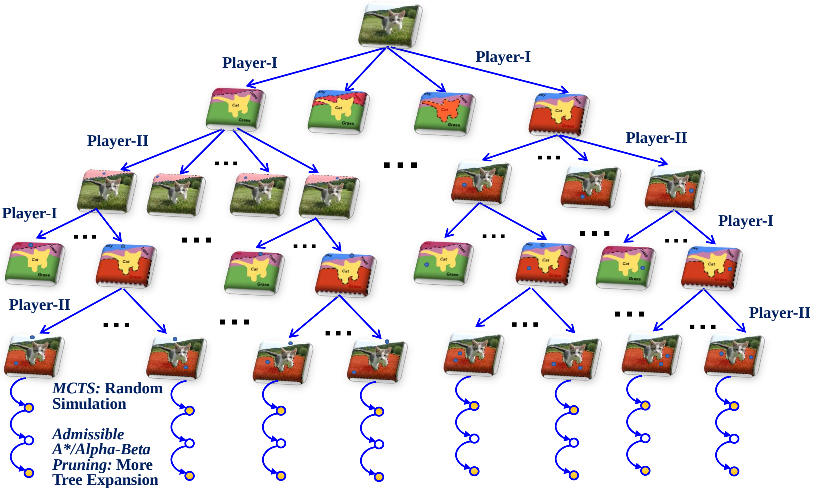
Theorem

Let N be a Lipschitz network with a Lipschitz constant $\hbar_c$ for every class $c \in C$. If

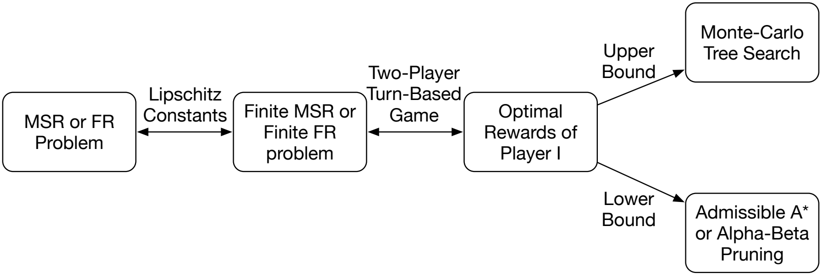
$$d(k, \tau) \leq \frac{2g(\alpha', N(\alpha'))}{\max_{c' \in C, c' \neq N(\alpha')}(\hbar_{N(\alpha')} + \hbar_{c'})}$$

for all τ-grid inputs $\alpha' \in G(\alpha, k, d)$, then we can use $\texttt{FMSR}(\tau, k, d, \alpha, c)$ *to estimate* $\texttt{MSR}(k, d, \alpha, c)$ *with an error bound* $\frac{1}{2}d(k, \tau)$.

Two Player Game

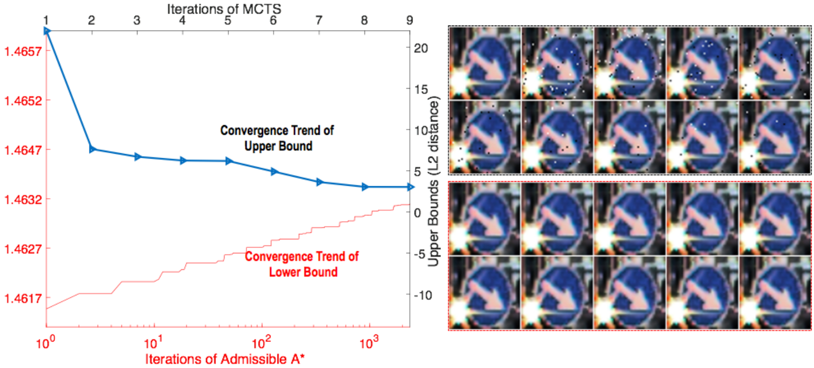


Flow of Reductions



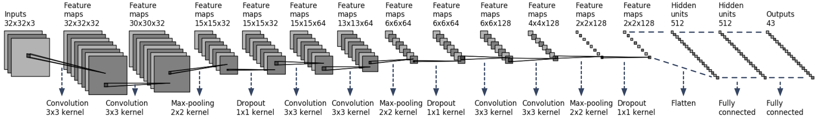
Outline

Convergence of Lower and Upper Bounds



Experimental Results: GTSRB

Image Classification Network for The German Traffic Sign Recognition Benchmark

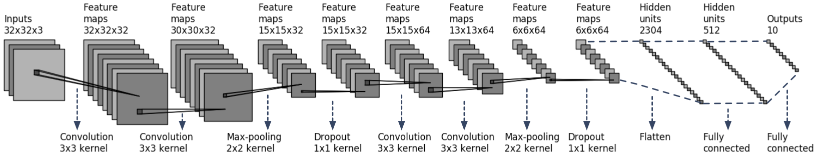


Total params: 571,723

Experimental Results: GTSRB

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

Comparison with Existing Tools on Finding Upper Bounds

L_0	MNIST				CIFAR10[1]			
	Distance		Time(s)		Distance		Time(s)	
	mean	std	mean	std	mean	std	mean	std
DeepGame	**6.11**	**2.48**	**4.06**	**1.62**	**2.86**	**1.97**	**5.12**	**3.62**
CW [1]	7.07	4.91	17.06	1.80	3.52	2.67	15.61	5.84
L0-TRE [5]	10.85	6.15	0.17	0.06	2.62	2.55	0.25	0.05
DLV [2]	13.02	5.34	180.79	64.01	3.52	2.23	157.72	21.09
SafeCV [6]	27.96	17.77	12.37	7.71	9.19	9.42	26.31	78.38
JSMA [3]	33.86	22.07	3.16	2.62	19.61	20.94	0.79	1.15

Comparison with Existing Tools on Finding Upper Bounds

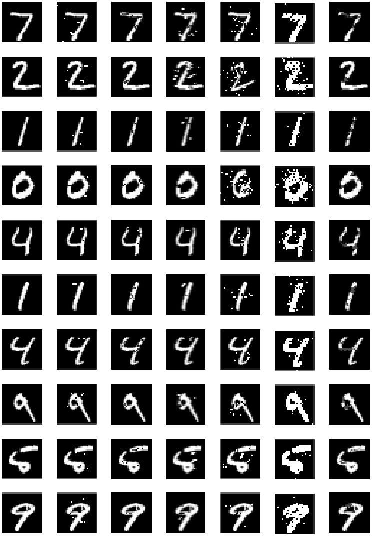


Figure: ‘original’, ‘DeepGame’, ‘CW’, ‘L0-TRE’, ‘DLV’, ‘SafeCV’, ‘JSMA’.

Comparison with Existing Tools on Finding Upper Bounds

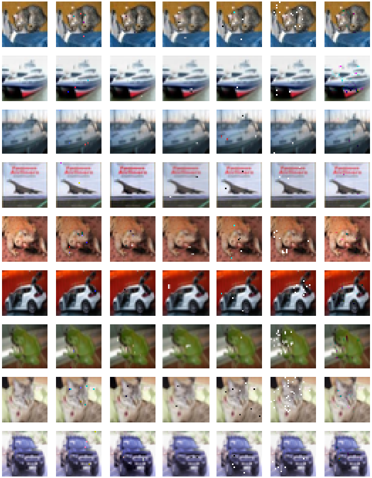


Figure: 'original', 'DeepGame', 'CW', 'L0-TRE', 'DLV', 'SafeCV', 'JSMA'.

Nexar Traffic Challenge

(a) (b) (c)

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

Conclusions and Future Works

- Pointwise Robustness (this talk)
- Network Robustness
- or more fundamentally, Lipschitz continuity, mutual information, etc
- model interpretability

Reference

Nicholas Carlini and David A. Wagner.
Towards evaluating the robustness of neural networks.
CoRR, abs/1608.04644, 2016.

Xiaowei Huang, Marta Kwiatkowska, Sen Wang, and Min Wu.
Safety verification of deep neural networks.
In *CAV 2017*, pages 3–29, 2017.

Nicolas Papernot, Patrick D. McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami.
The limitations of deep learning in adversarial settings.
CoRR, abs/1511.07528, 2015.

Wenjie Ruan, Xiaowei Huang, and Marta Kwiatkowska.
Reachability analysis of deep neural networks with provable guarantees.
In *IJCAI-2018*, 2018.

Wenjie Ruan, Min Wu, Youcheng Sun, Xiaowei Huang, Daniel Kroening, and Marta Kwiatkowska.
Global robustness evaluation of deep neural networks with provable guarantees for L0 norm.
CoRR, abs/1804.05805, 2018.

Matthew Wicker, Xiaowei Huang, and Marta Kwiatkowska.
Feature-guided black-box safety testing of deep neural networks.
In *TACAS 2018*, 2018.