A Dual Approach to Scalable Verification of Deep Networks

Krishnamurthy (Dj) Dvijotham

DeepMind

How Artificial Intelligence Is Making Energy Smarter and Cleaner

'It's going to create a revolution': how AI is transforming the NHS

How banks and finance firms are using AI to better engage with and understand you

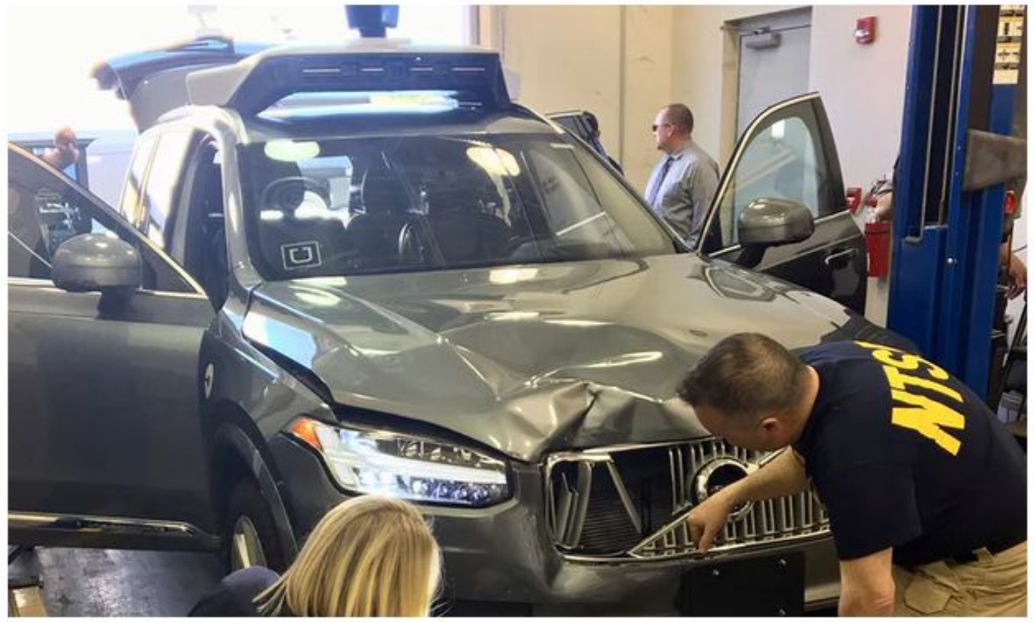
AI is a powerful technology ...

... with power comes responsibility

AI systems in the wild

Arizona suspends Uber's self-driving car testing after fatality

Governor Doug Ducey tells Uber crash raises concerns about its ability to safely test technology



Is AI a threat to fair lending?

Amazon Echo nightmare: private conversation sent to contact

Couple learns of recording after husband's employee calls about receiving audio files

Need strong safety checks on AI systems

Supervised learning

Specification

(training data)

Implementation
(Predictor)

Learning

(neural networks)
(decision trees) ..

Training-data specifications not enough

US opens investigation into Tesla after fatal crash

Dave Lee
North America technology reporter

.@TeslaMotors Model S autopilot camera misreads 101 sign as 105 speed limit at 87/101 junction San Jose. Reproduced every day this week.

8:40 PM - 14 Jul 2017

Researchers Find a Malicious Way to Meddle with Autonomous Cars

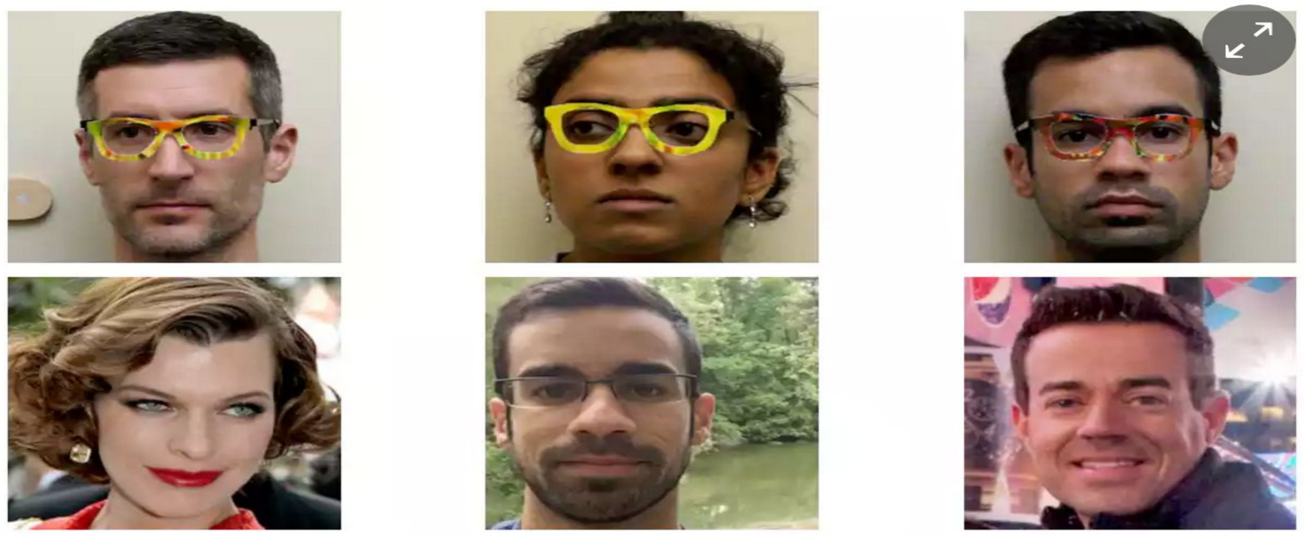
Robust Physical-World Attacks on Machine Learning Models

Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, Dawn Song

(Submitted on 27 Jul 2017 (v1), last revised 30 Jul 2017 (this version, v2))

Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition

Eyewear printed with a wild pattern can be enough to fool commercial systems into misidentification, research shows



Impact on bias and fairness

WIRED — Courts Are Using AI to Sentence Criminals. That Must Stop Now

JASON TASHEA OPINION 04.17.17 07:00 AM

SHARE

COURTS ARE USING AI TO SENTENCE CRIMINALS. THAT

We need richer specifications for ML models

- robustness to adversaries
- fairness and unbiasedness
- Physics-compliant (satisfies conservation of energy, conservation of momentum etc.)
- ....

adversarial examples as a case-study

Specification-driven ML

Verification

(MIP, SMT, duality etc.)

Specification

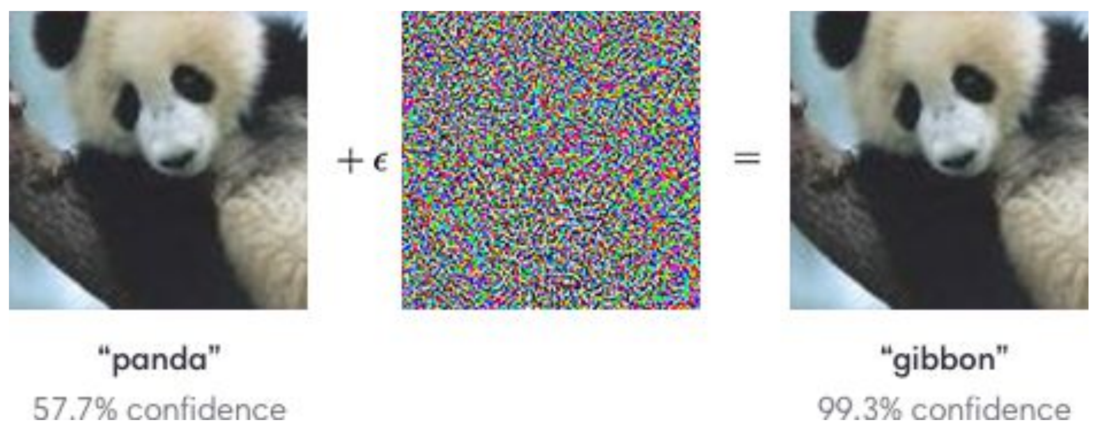
(training data)
+(robustness to adv examples)
+(consistent with physics)
+ (respect label semantics) ...

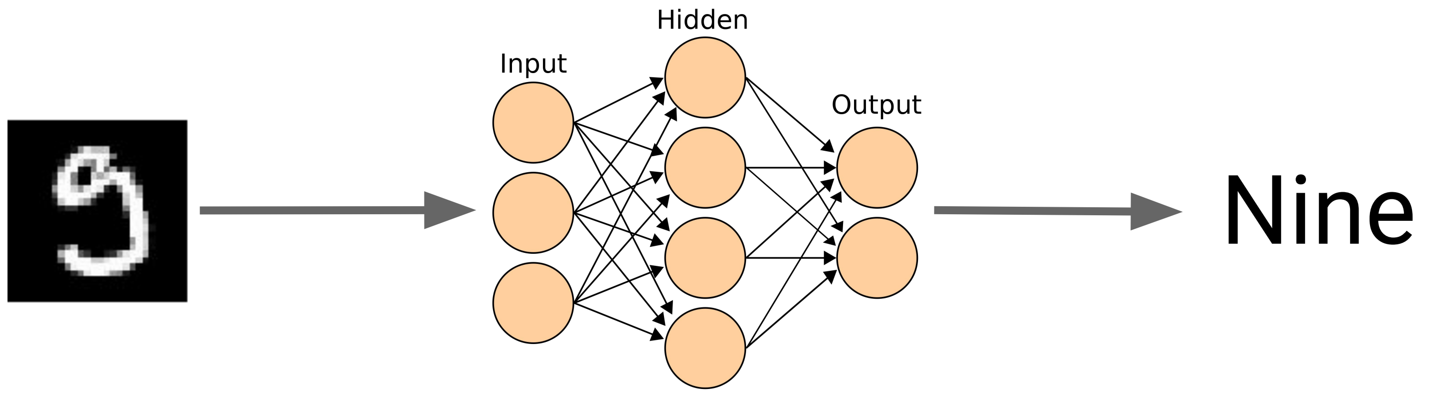
Verified Implementation
(Predictor + Checker)

Learning

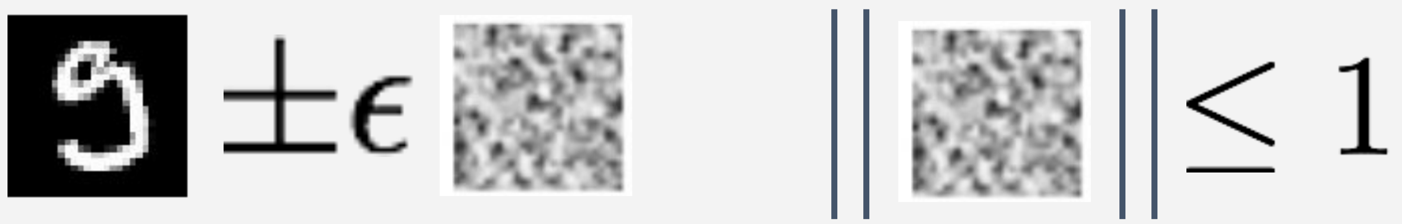
architectures (NNs, RF, ...)
verifiers (NNs, MIPs, LPs, ...)
verified training (PVT, ...)

Adversarial attacks on image classifiers

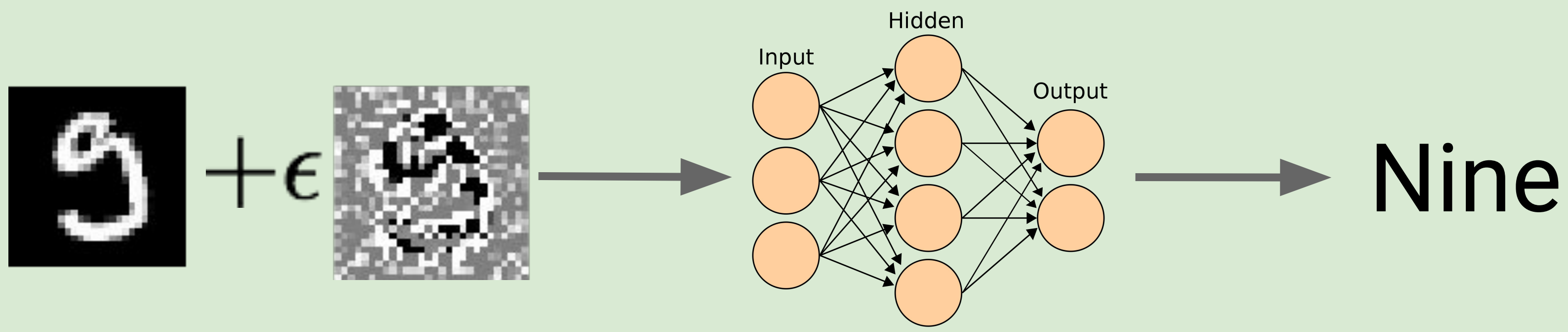




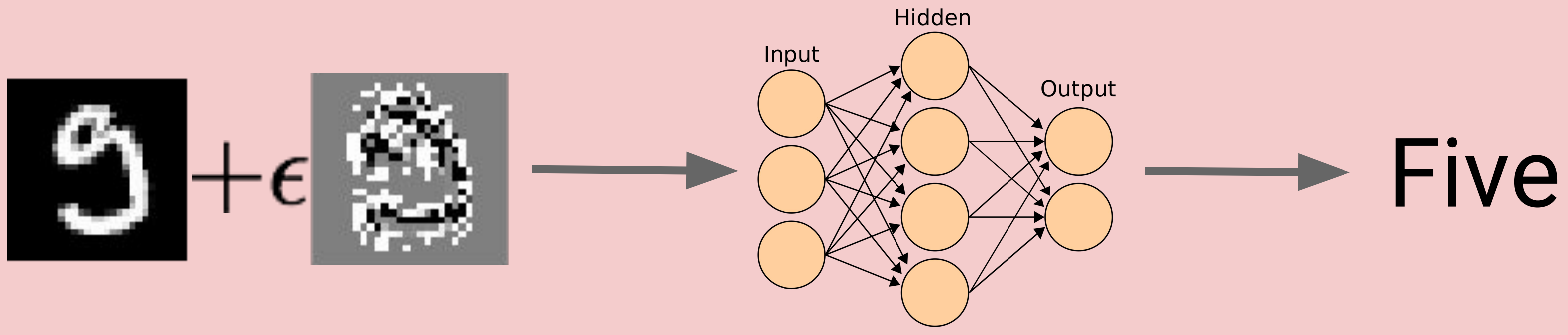
Specification: Output remains "Nine" for **ALL IMAGES** of the form



Projected Gradient Attack



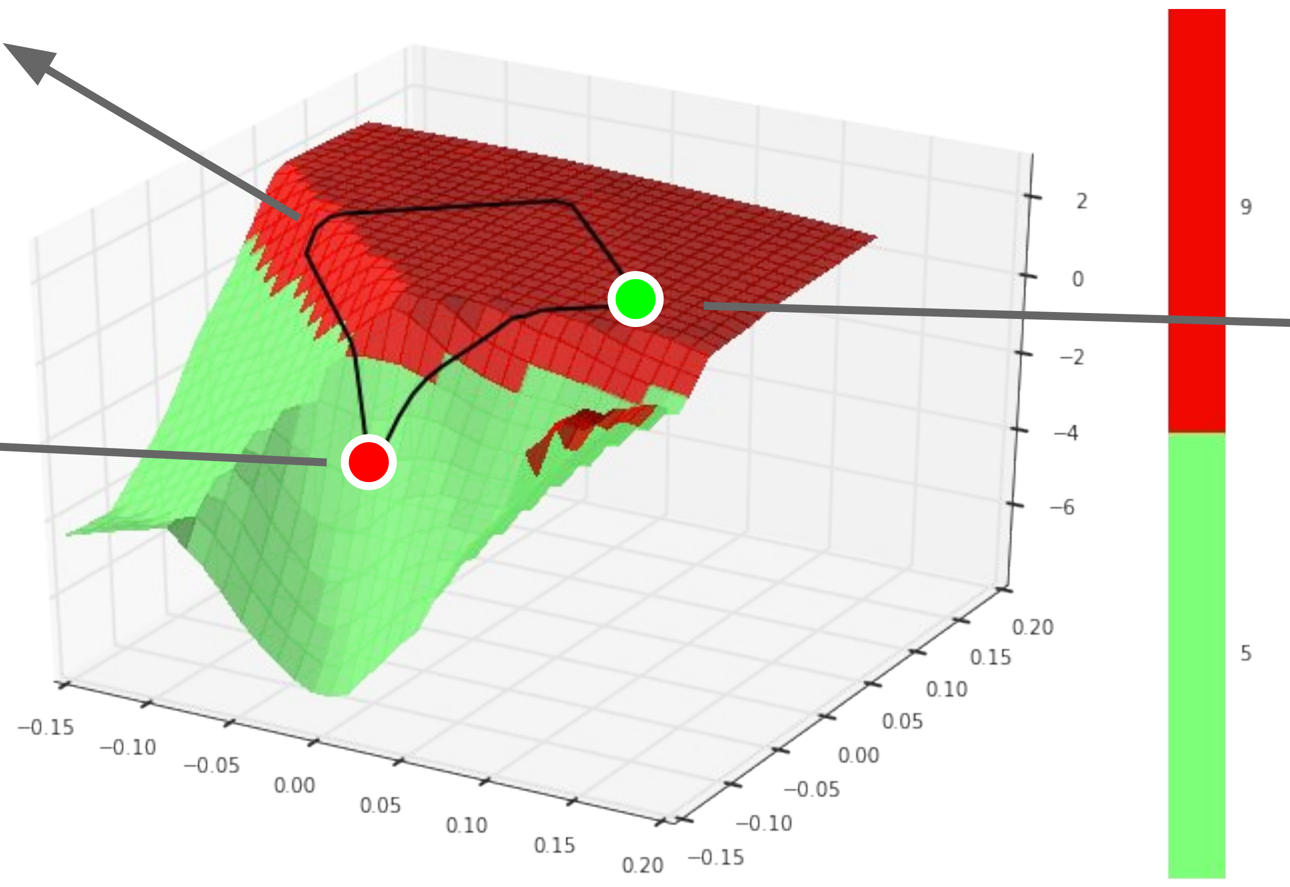
True worst case



Why PGD attack fails?

ε-ball

Worst-case adversarial attack



Projected Gradient attack

Need verification: Provable guarantee that no adversarial attack can succeed

Defense strategies don't really work

Evaluation of NIPS competition winners/published papers

- Non-differentiable models (ICLR 2018)
- Generative-denoising (ICLR 2018)
- Denoising with semantic features (NIPS Competition winner)
- Constraining input gradients (ICML 2017)
- Stochasticity / Ensembling (ICLR 2018, NIPS 2nd place)

Defense Strategy	Standardized Evaluation
CIFAR-10 (e = 8)	
Non-differentiability	43%
Generative modeling	46%
Adversarial Training	45%
ImageNet (e = 2)	
Stochasticity	32%
Denoising	61%

DeepMind

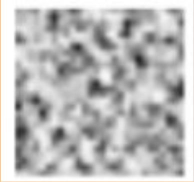
Athalye et al. *Gradient obfuscation* ... ICML 2018

Uesato et al. *Dangers of weak attacks.* ICML 2018

Hardness of verification in general

Verification by enumeration:

Discretize space of perturbations

 $\pm\epsilon$ 

(Perturbation size) $^{(\#\text{Pixels})}$ - search space grows exponentially!

- Verifying 10% perturbation attack on MNIST takes $O(10^{1000})$ CPU-years
- NP-hard to find constant factor approx of optimal attack [Weng et al, 2018]

Need to trade-off scalability and completeness of verification procedure

Sound and complete verification algorithms

Intelligent Brute-Force Search

Guy Katz, Clark Barrett, David Dill, Kyle Julian, Mykel Kochenderfer. *Reluplex: An efficient SMT solver for verifying deep neural networks*. International Conference on Computer Aided Verification. 2017. [PDF]

Formal Verification of Piece-Wise Linear Feed-Forward Neural Networks

Ruediger Ehlers

Satisfiability Modulo Theory

Piecewise Linear Neural Network Verification: A comparative Study

Rudy Bunel, Ilker Turksaslan, Philip H.S. Torr, Pushmeet Kohli, M. Pawan Kumar

Evaluating Robustness of Neural Networks with Mixed Integer Programming

Vincent Tjeng, Kai Xiao, Russ Tedrake

Mixed-Integer Programming

Encouraging progress but limited scalability

Incomplete verification algorithms

Partial search on abstraction/relaxation

Provable defenses against adversarial examples via the convex outer adversarial polytope
E Wong, Z Kolter - International Conference on Machine ..., 2018 - proceedings.mlr.press

Certified defenses against adversarial examples
A Raghunathan, J Steinhardt, P Liang - arXiv preprint arXiv:1801.09344, 2018 - arxiv.org

- Use convex relaxation of nonlinearity
- LP, SDP, Convex program

Ai 2: Safety and robustness certification of neural networks with abstract interpretation
T Gehr, M Mirman, D Drachsler-Cohen... - Security and Privacy ..., 2018 - computer.org

Towards Fast Computation of **Certified** Robustness for ReLU Networks
TW Weng, H Zhang, H Chen, Z Song, CJ Hsieh... - arXiv preprint arXiv ..., 2018 - arxiv.org

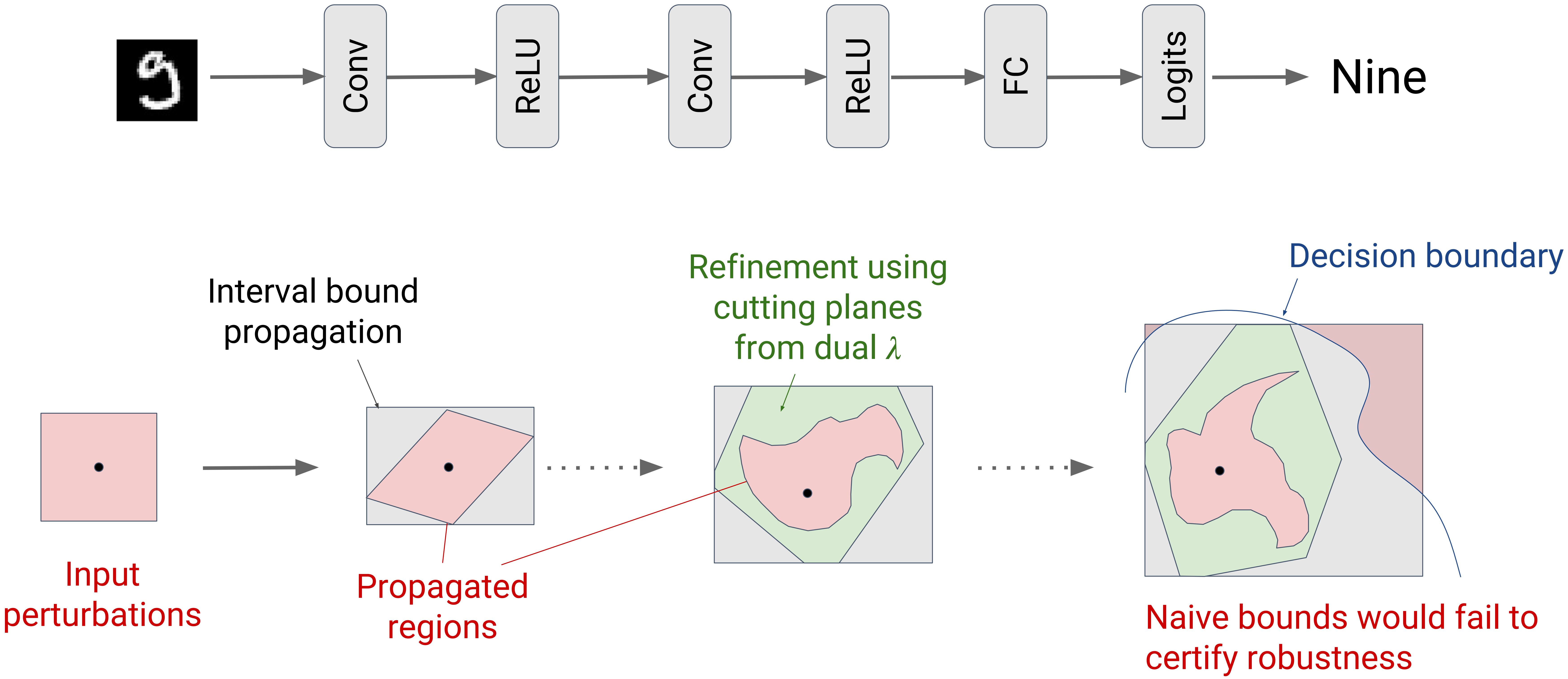
- Use abstraction of nonlinearity
- Propagate “simple” abstractions
- Symbolic bounds, zonotopes etc

Scalable but limited generality/completeness

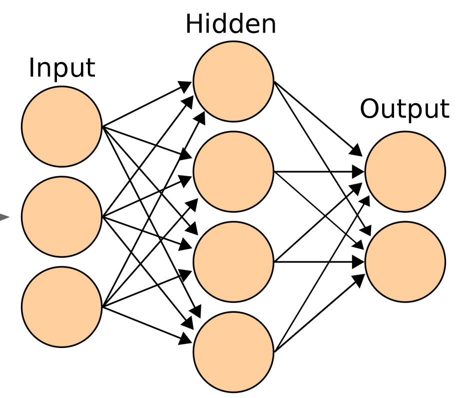
Comparison of approaches

	Completeness	Complexity	Backprop-friendly	Handles non piecewise-linear
Reluplex	✓	?	✗	✗
Bunel 17	✓	?	✗	✗
AI2	✗	✓?	✓	✗
Kolter/Wong 18	✗	✓?	✓	✗
Raghunathan 18	✗	✓?	✓	Only single hidden layer
This paper	✗	✓	✓	✓

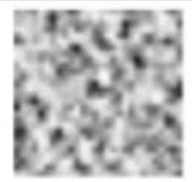
Verification process geometric view



Formulation of verification

x_{in}  x_{out}

$$\forall x_{in} \in \mathcal{S} \quad c^T x_{out} + d \leq 0$$

$\forall x_{in} \in$  $\pm\epsilon$  $\quad x_{out;5} - x_{out;9} \leq 0$

Formulation of verification

$$\begin{aligned} \max \quad & c^T x_K + d \\ \text{Subject to } & x_{k+1} = h_k(x_k) \quad k = 0, \ldots, K-1 \\ & x_0 \in \mathcal{S} \end{aligned}$$

$$\begin{aligned} \max \quad & c^T x_K + d + \sum_{k=0}^{K-1} \lambda_k^T \left(x_{k+1} - h_k(x_k)\right) \\ \text{Subject to } & x_0 \in \mathcal{S} \\ & \boxed{x_k \in [l_k, u_k]} \end{aligned}$$

From interval arithmetic

$$\sum_k \boxed{\max_{x_k \in [l_k, u_k]} \lambda_k^T x_k - \lambda_{k+1}^T h_k(x_k)} + d$$

Solved analytically for most common h

DeepMind

Verification as optimization

$$f(\boldsymbol{\lambda}) = \sum_k \max_{x_k \in [l_k, u_k]} \lambda_k^T x_k - \lambda_{k+1}^T h_k(x_k) + d$$

For any choice of $\boldsymbol{\lambda}$,

$$\begin{aligned} \max \quad & c^T x_K + d \\ \text{Subject to} \quad & x_{k+1} = h_k(x_k) \quad k = 0, \ldots, K-1 \\ & x_0 \in \mathcal{S} \end{aligned} \qquad \leq \qquad f(\boldsymbol{\lambda})$$

By weak duality

Obtain best possible bound by solving $\min_{\boldsymbol{\lambda}} f(\boldsymbol{\lambda})$

Unconstrained convex program

Theoretical results

Can verification be done tractably under special assumptions?

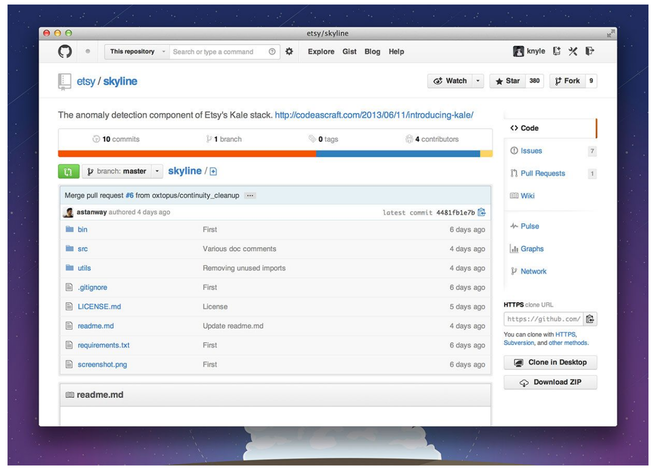
Assumptions

2-norm constraint on input $\|x - x'\|_2 \leq \epsilon$, single hidden layer

Theorem

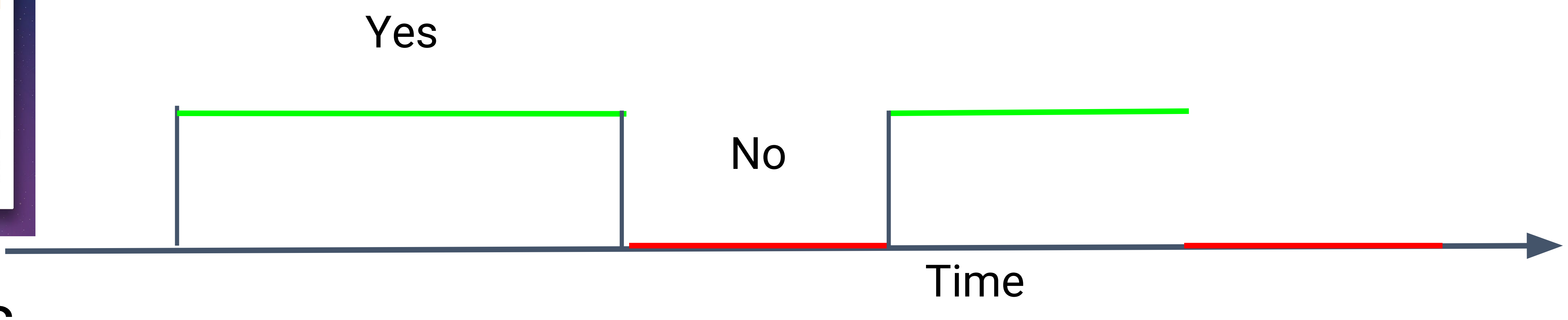
1. If $\epsilon < \kappa(NN)$, can solve verification problem exactly using projected gradient algorithm
2. Otherwise, can obtain $\zeta(NN)\epsilon^3$ additive approximation by solving a trust region problem

Results: Classifier stability

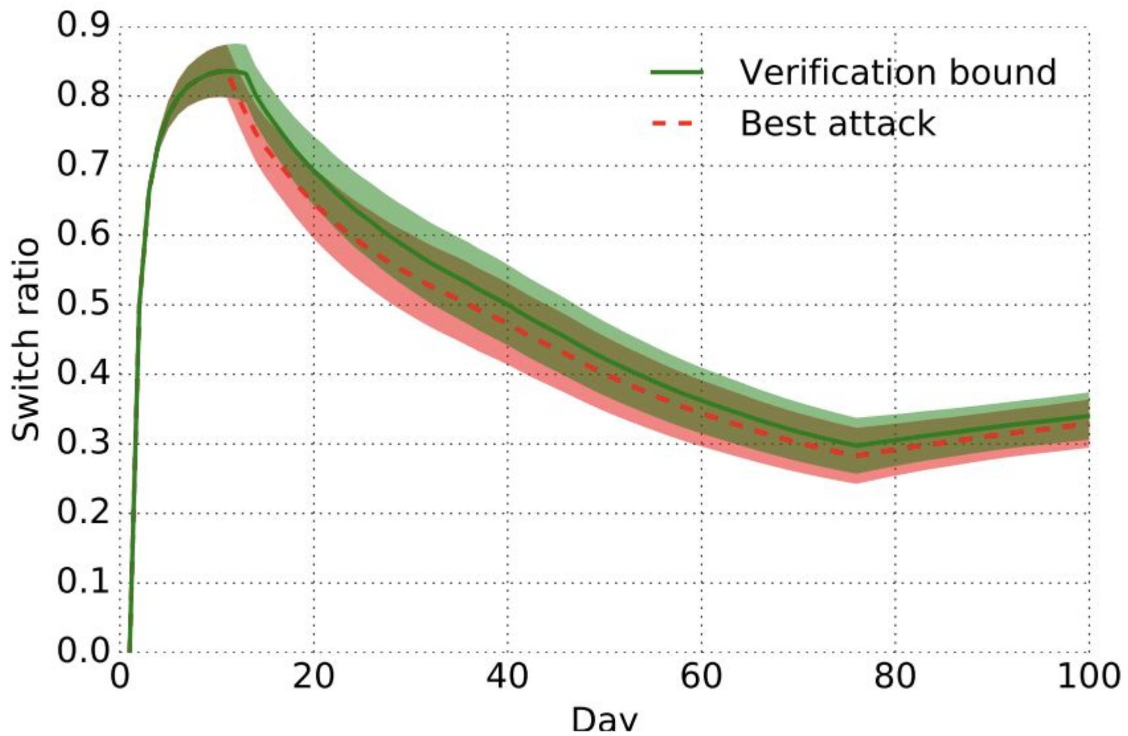


# of commits

# of commiters

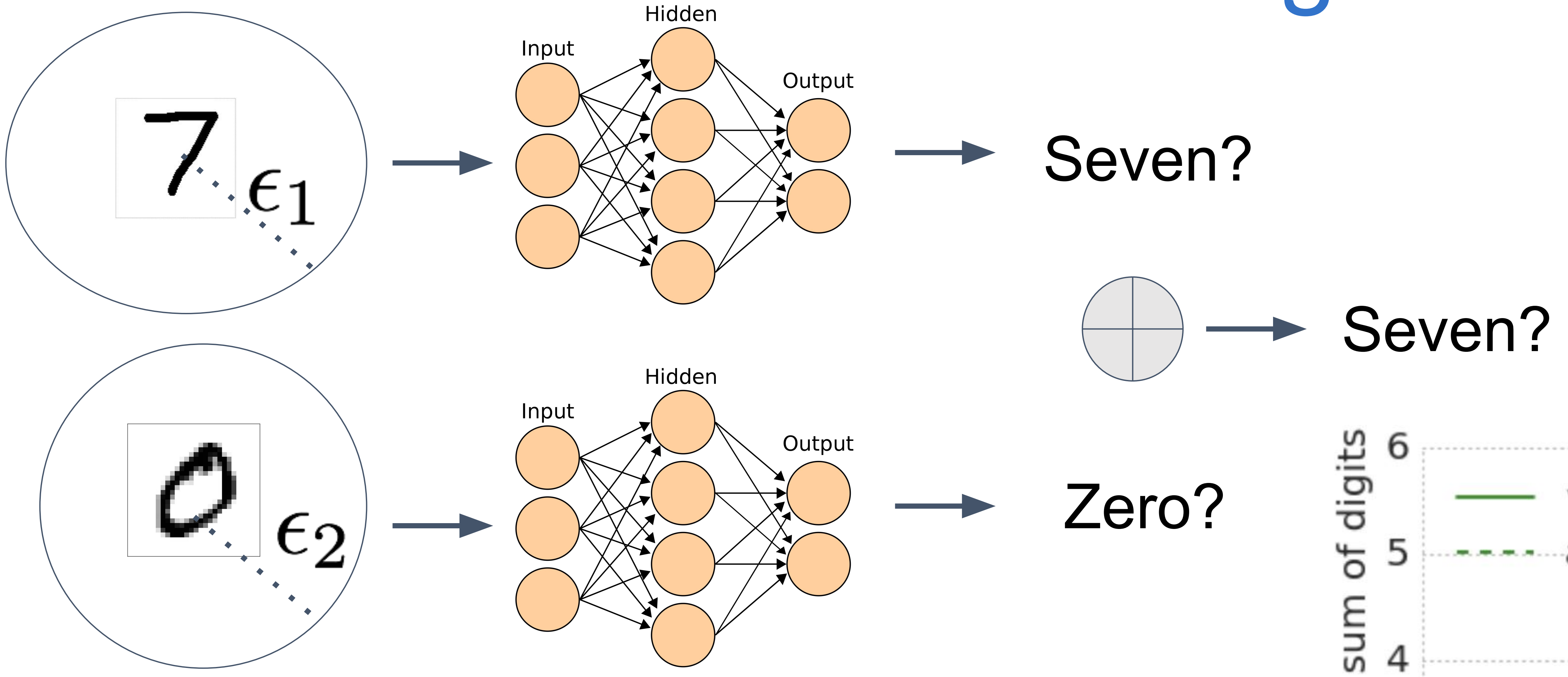


How often does the prediction switch as features evolve?



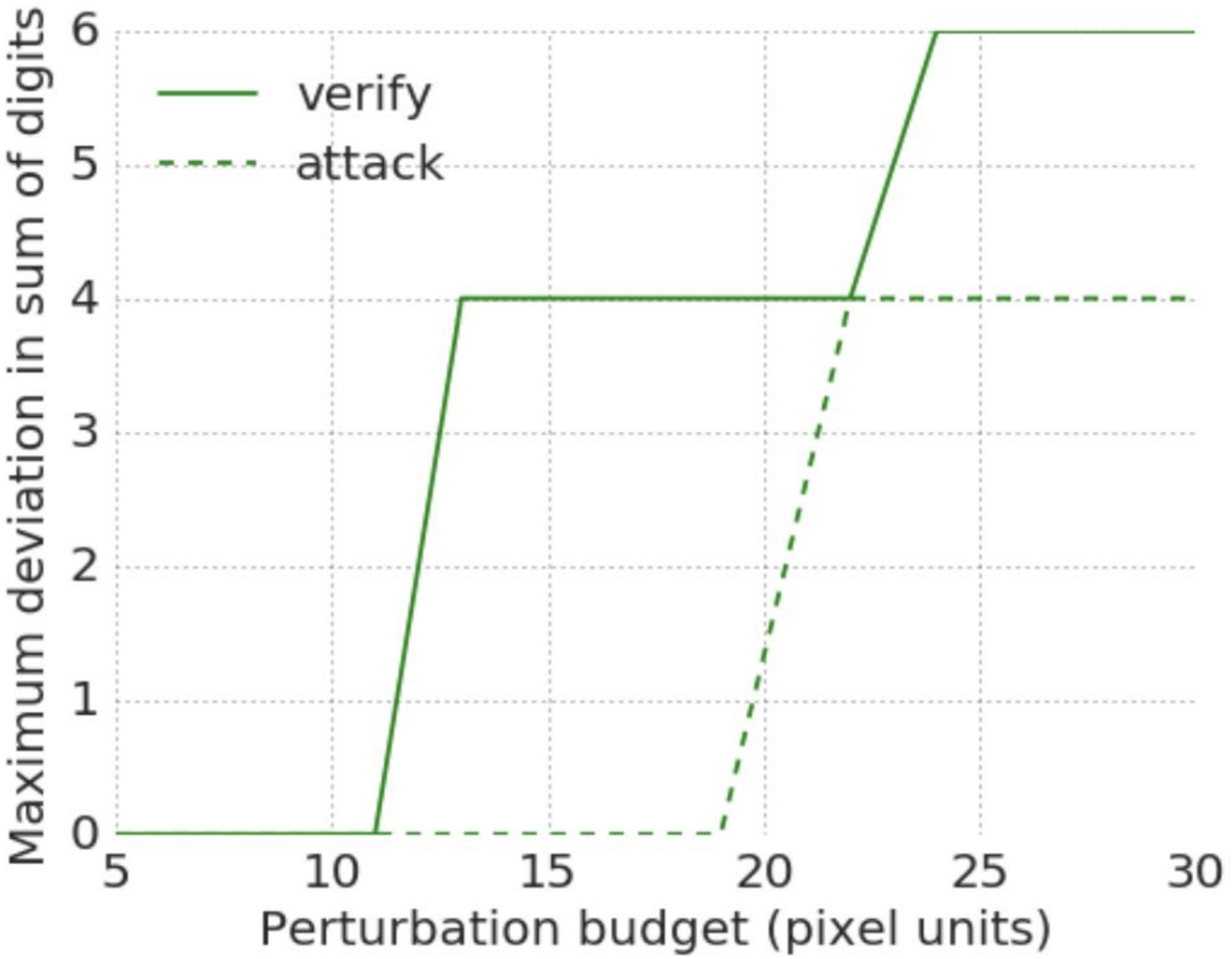
Bounds on switching frequency: best attack vs verified bound, averaged over several datasets (github repositories data)

Results: Digit Sum



How much can the sum of predictions differ from true sum (7) given budget

$$\epsilon_1 + \epsilon_2 \leq \epsilon$$

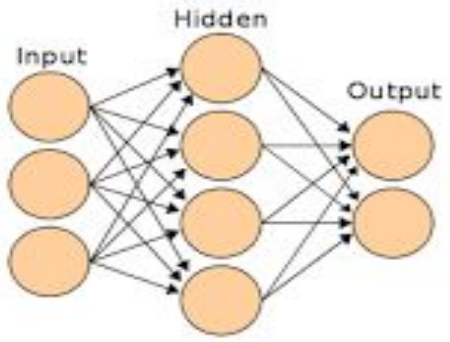


Learning verifiers

Verified training

- Networks are not robust by construction
- Standard techniques for enhancing robustness fail (rh/4537)
- Some networks are "easy to verify" (rh/4343)
- Solving dual optimization for each training example is a huge overhead

Learn a verifier to "guess" the right cutting planes

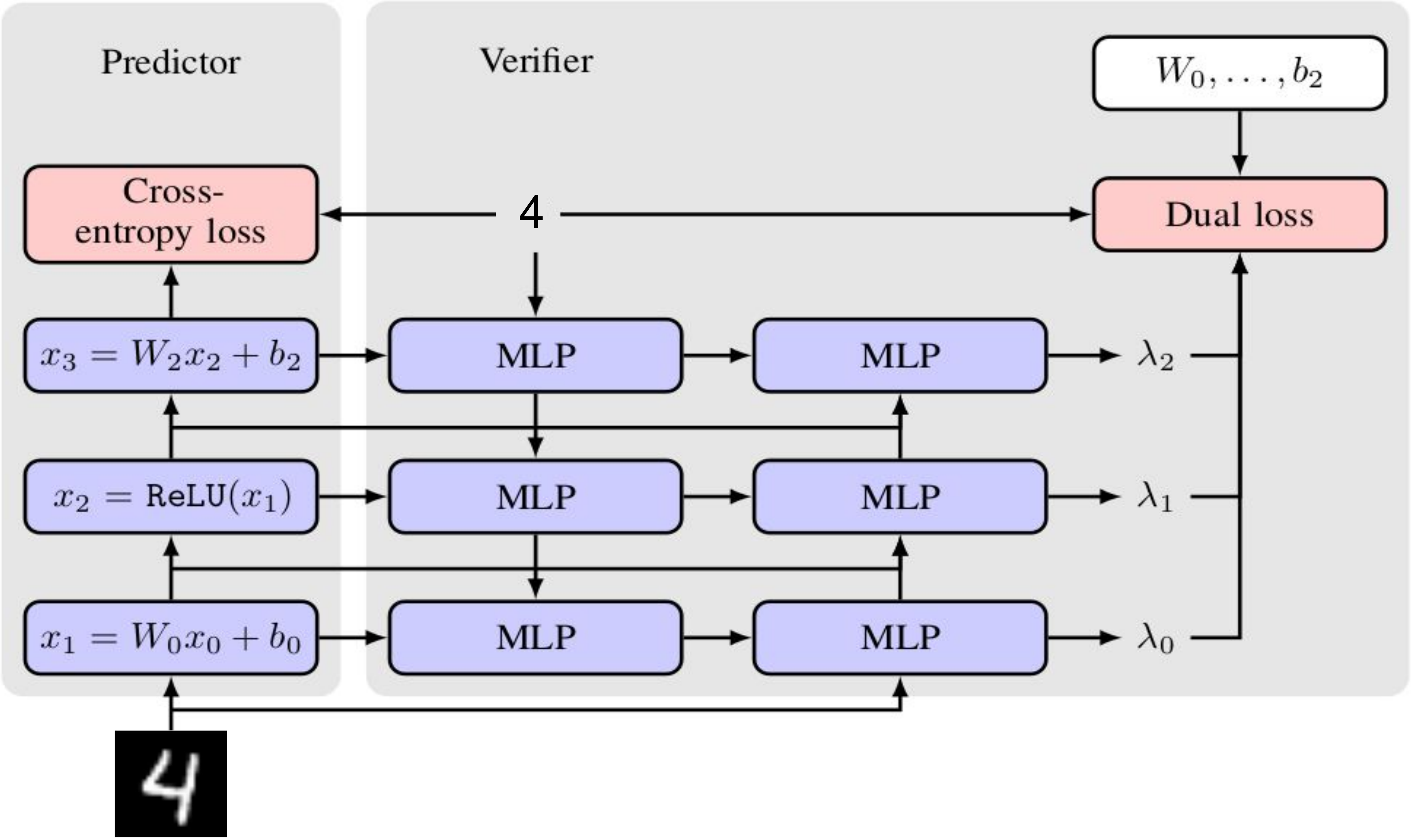


λ

Bound on adversarial error rate

Verifier network

Predictor verifier training



min *cross_entropy_loss* +
κ * *dual_loss*

w.r.t. *predictor_weights*
verifier_weights

Results

Problem	Method	Perturbation size (pixel values)	Nominal Error	PGD Attack Error	Verified error
MNIST	Baseline	25 / 255	0.77%	52.94%	100.00%
MNIST	Wong and Kolter [1]		1.80%	4.11%	5.82%
MNIST	Wong et al. [4]*		1.26%	—	4.48%
MNIST	Madry et al. [2]		**0.60%**	4.66%	100.00%
MNIST	Predictor-Verifier		1.01%	**3.16%**	**4.21%**
SVHN	Baseline	3 / 255	**6.57%**	87.45%	100.00%
SVHN	Wong and Kolter [1]		20.38%	33.74%	40.67%
SVHN	Madry et al. [2]		7.04%	**23.63%**	100.00%
SVHN	Predictor-Verifier		16.29%	33.14%	**37.56%**
CIFAR-10	Baseline	8 / 255	**26.37%**	99.99%	100.00%
CIFAR-10	Madry et al. [2]		39.00%	68.08%	100.00%
CIFAR-10	Wong et al. [4]*		72.24%	—	79.25%
CIFAR-10	Predictor-Verifier		51.35%	**67.28%**	**73.01%**

Results

Problem	Method	Perturbation size (pixel values)	Nominal Error	PGD Attack Error	Verified error
MNIST	Wong et al. [4]*	25 / 255	3.13%	—	3.13%
MNIST	Lamb et al. [5]		—	1.91%	—
MNIST	Predictor-Verifier		0.93%	**1.79%**	4.41%
MNIST	Predictor-Verifier		1.01%	2.43%	**2.60%**
CIFAR-10	Madry et al. [2]	8 / 255	12.70%	**54.20%**	—
CIFAR-10	Wong et al. [4]*		70.77%	—	**70.95%**
CIFAR-10	Predictor-Verifier		51.35%	67.28%	73.01%
CIFAR-10	Predictor-Verifier		56.67%	—	71.35%

Uses cascaded ensemble

Future outlook

Specification-driven ML

Verification

(MIP, SMT, duality etc.)

Specification

(training data)
+(robustness to adv examples)
+(consistent with physics)
+ (respect label semantics) ...

Verified Implementation
(Predictor + Checker)

Learning

architectures (NNs, RF, ...)
verifiers (NNs, MIPs, LPs, ...)
verified training

Open questions and challenges

Tractable verification: Under what conditions can verification be done tractably? Results for single hidden layer networks - can be extended beyond?

Theoretical foundations: Integrating learning into verification leads to easily verifiable networks with small duality gap. Can this be explained theoretically?

Reinforcement learning guided verification: Can we use RL inside the search process of a verification algorithm to guide the search?

Fundamental tradeoffs: If we are trying to verify multiple graded specifications, can we quantify fundamental tradeoffs? Nominal performance vs robustness?

Questions?

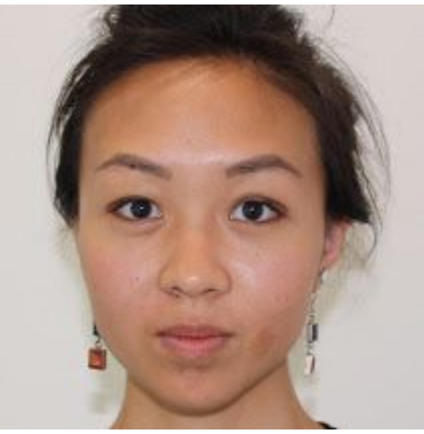
Sven Gowal

Robert Stanforth

Timothy Mann

Pushmeet Kohli

Rudy Bunel



Chongli Qin

https://arxiv.org/abs/1803.06567
https://arxiv.org/abs/1805.10265