Holistic Adversarial Robustness for Deep Learning

Pin-Yu Chen

IBM Research AI

pin-yu.chen@ibm.com

www.pinyuchen.com @pinyuchenTW

Machine Learning Summer School (MLSS@Taipei)

August 2021

IBM Research

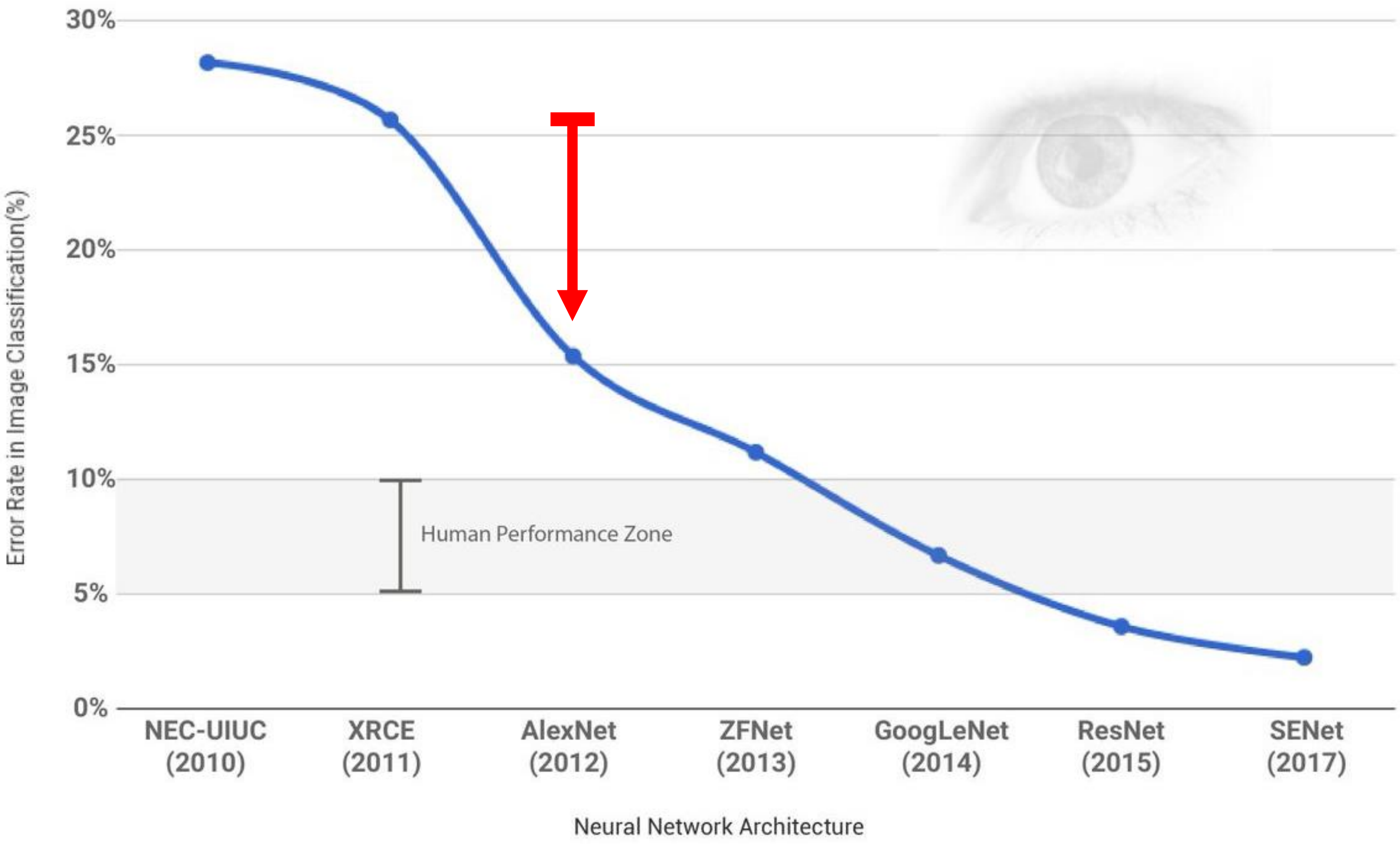
Outline

- First Part:
  1. Introduction
  2. Adversarial Attack
  3. Applications and Extensions
  4. Q&A
- Second Part:
  1. Model Reprogramming
  2. Defense
  3. Verification
  4. Conclusion
  5. Resources
  6. Q&A

#ImageNet Generation

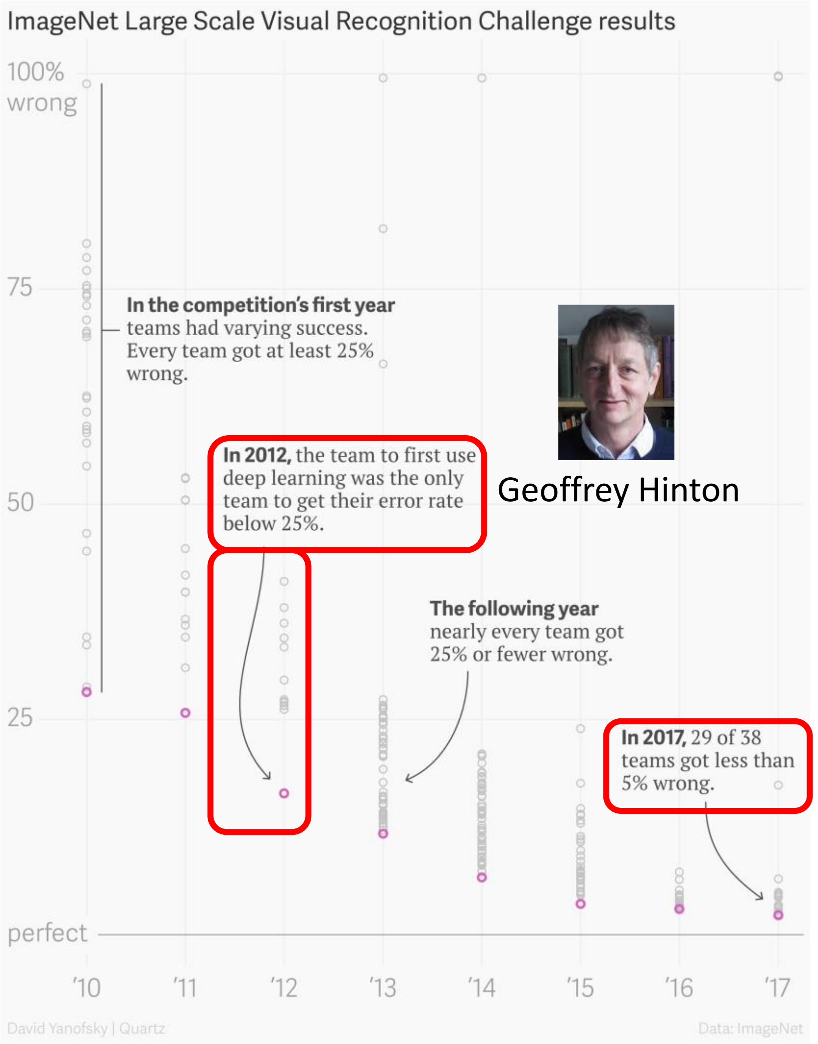
https://medium.com/syncedreview/sensetime-trains-imagenet-alexnet-in-record-1-5-minutes-e944ab049b2c

ImageNet Challenges

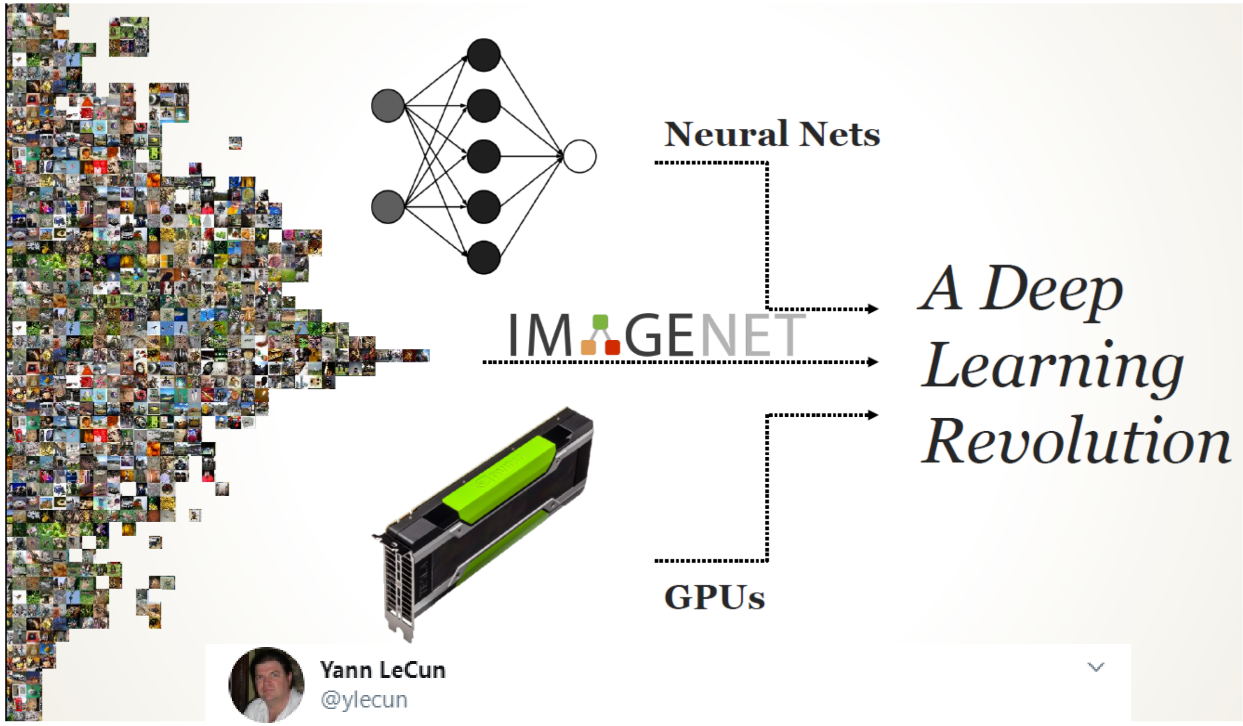


https://towardsdatascience.com/transfer-learning-in-tensorflow-9e4f7eae3bb4

The Deep Learning Revolution. What's next?



What's Next?



Yann LeCun
@ylecun

Replying to @ylecun @GaryMarcus and @titudeadjust

DL is not an "algorithm". It's merely the concept of building a machine by assembling parameterized functional blocks and training them with some sort of gradient-based optimization method. That's it.
You are free to choose your architecture, learning paradigm, prior, etc...1/2

http://image-net.org/challenges/talks_2017/imagenet_ilsvrc2017_v1.0.pdf

https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world/

What happens when you do well on ImageNet?

The gap between AI development and deployment

How we develop AI

How we deploy AI

AI revolution is coming, but *Are We Prepared ?*

- According to a recent Gartner report, 30% of cyberattacks by 2022 will involve data poisoning, model theft or adversarial examples.
- However, industry is underprepared. In a survey of 28 organizations spanning small as well as large organizations, 25 organizations did not know how to secure their AI systems.

DEFENSE

Pentagon actively working to combat adversarial AI

The Great Adversarial Examples

ostrich | safe | shoe shop | vacuum

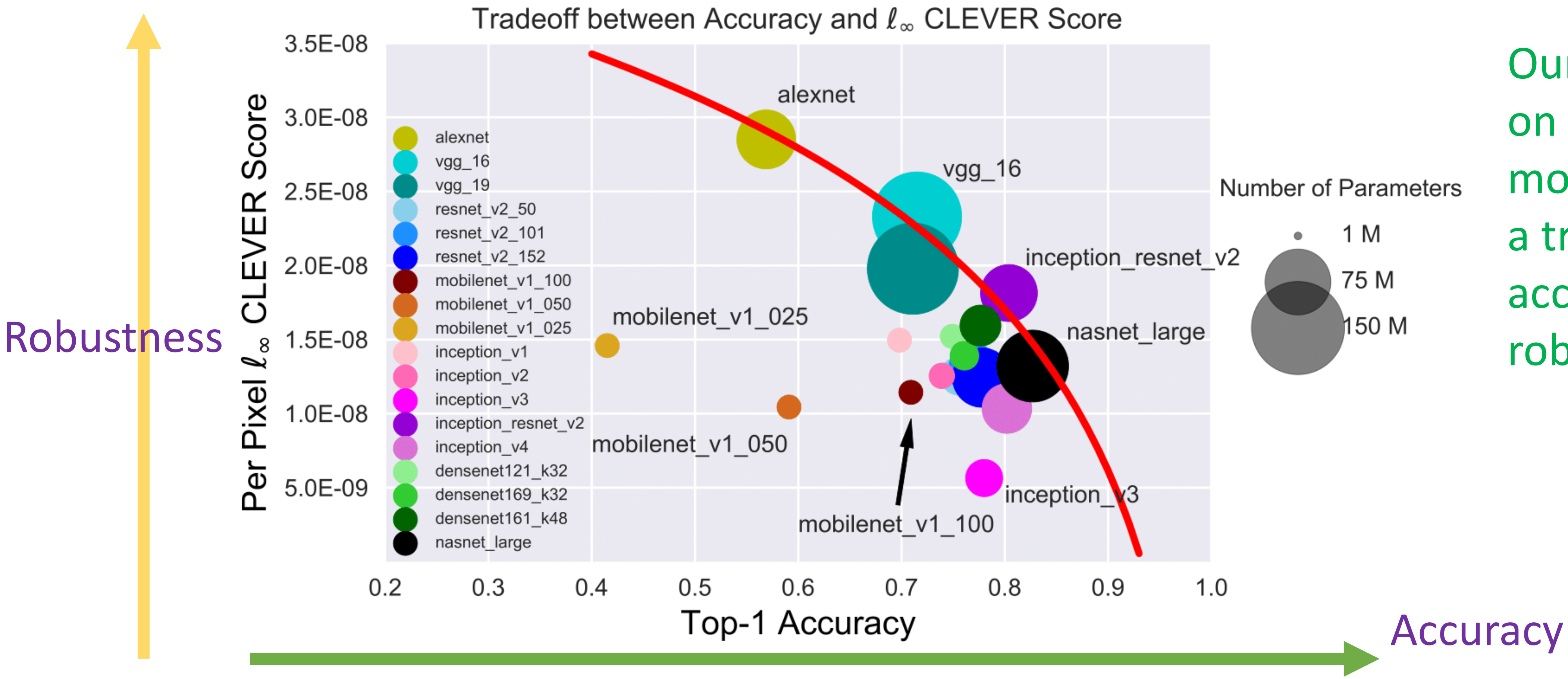
What is wrong with this AI model?

- This model is one of the BEST image classifier using neural networks
- Images and neural network models are NOT the only victims

EAD: Elastic-Net Attacks to Deep Neural Networks via Adversarial Examples, P.-Y. Chen*, Y. Sharma*, H. Zhang, J. Yi, and C.-J. Hsieh, AAAI 2018

Accuracy ≠ Adversarial Robustness

- Solely pursuing for high-accuracy AI model may get us in trouble…



Our benchmark on 18 ImageNet models reveals a tradeoff in accuracy and robustness

Is Robustness the Cost of Accuracy? A Comprehensive Study on the Robustness of 18 Deep Image Classification Models, Dong Su*, Huan Zhang*, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao, ECCV 2018

Adversarial examples: the evil doublegangers

NOT
A HERO

source: Google Images

Why adversarial (worst-case) robustness matters?

- Prevent <u>prediction-evasive</u> manipulation on deployed models

Build trust in AI: address inconsistent decision making between humans and machines & misinformation

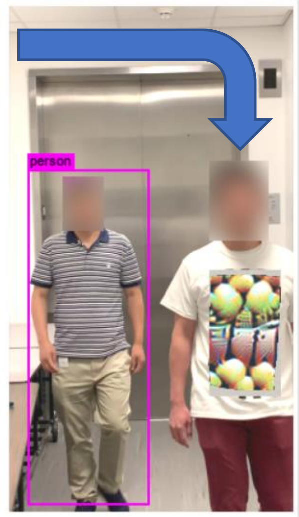
Assess negative impacts in high-stakes, safety-critical tasks

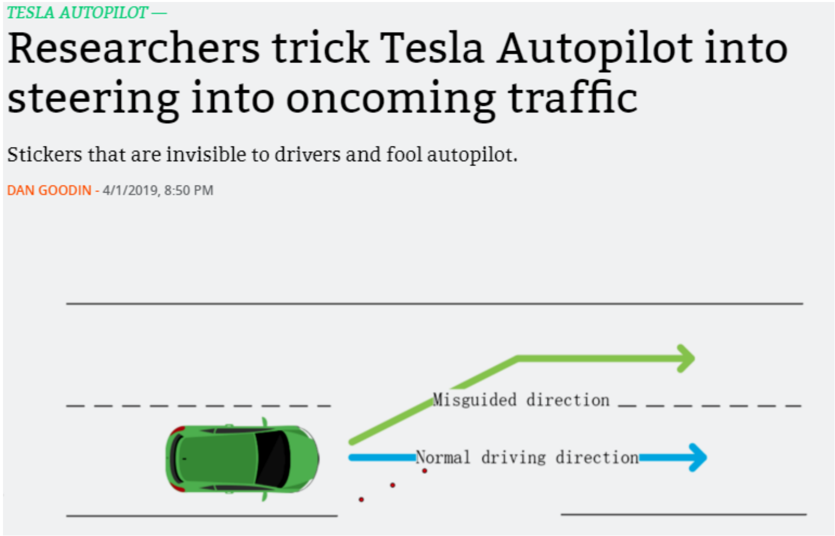
Understand limitation in current machine learning methods

Prevent loss in revenue and reputation

Ensure safe and responsible use in AI

Adversarial T-shirt




TESLA AUTOPILOT —

Researchers trick Tesla Autopilot into steering into oncoming traffic

Stickers that are invisible to drivers and fool autopilot.

DAN GOODIN - 4/1/2019, 8:50 PM


The Washington Post

Democracy Dies in Darkness

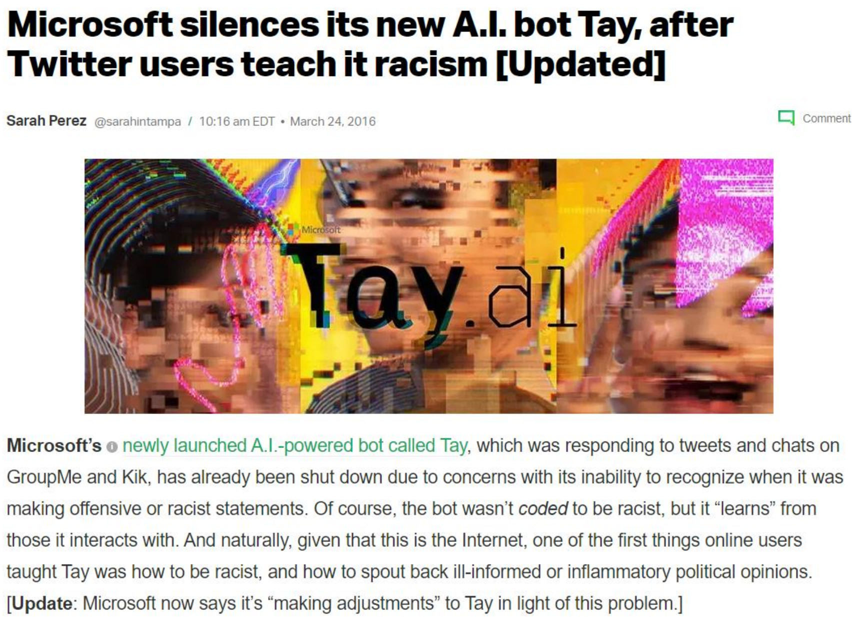
WorldViews

Syrian hackers claim AP hack that tipped stock market $136 billion. Is it terrorism?

This chart shows the Dow Jones Industrial Average during Tuesday afternoon's drop, caused by a fake A.P. tweet, inset at left.

By Max Fisher

April 23, 2013 at 4:31 p.m. EDT


Microsoft silences its new A.I. bot Tay, after Twitter users teach it racism [Updated]

Sarah Perez @sarahintampa / 10:16 am EDT • March 24, 2016

Comment

Microsoft's newly launched A.I.-powered bot called Tay, which was responding to tweets and chats on GroupMe and Kik, has already been shut down due to concerns with its inability to recognize when it was making offensive or racist statements. Of course, the bot wasn't *coded* to be racist, but it "learns" from those it interacts with. And naturally, given that this is the Internet, one of the first things online users taught Tay was how to be racist, and how to spout back ill-informed or inflammatory political opinions. [**Update**: Microsoft now says it's "making adjustments" to Tay in light of this problem.]

AI technology: Jewel of the Crown

Adversarial ML Threat Matrix

https://github.com/mitre/advmlthreatmatrix

Reconnaissance	Initial Access	Execution	Persistence	Model Evasion	Exfiltration	Impact
Acquire OSINT information: (Sub Techniques) 1. Arxiv 2. Public blogs 3. Press Releases 4. Conference Proceedings 5. Github Repository 6. Tweets	Pre-trained ML model with backdoor	Execute unsafe ML models (Sub Techniques) 1. ML models from compromised sources 2. Pickle embedding	Execute unsafe ML models (Sub Techniques) 1. ML models from compromised sources 2. Pickle embedding	Evasion Attack (Sub Techniques) 1. Offline Evasion 2. Online Evasion	Exfiltrate Training Data (Sub Techniques) 1. Membership inference attack 2. Model inversion	Defacement
ML Model Discovery (Sub Techniques) 1. Reveal ML model ontology – 2. Reveal ML model family –	Valid account	Execution via API	Account Manipulation		Model Stealing	Denial of Service
Gathering datasets	Phishing	Traditional Software attacks	Implant Container Image	Model Poisoning	Insecure Storage 1. Model File 2. Training data	Stolen Intellectual Property
Exploit physcial environment	External remote services			Data Poisoning (Sub Techniques) 1. Tainting data from acquisition – Label corruption 2. Tainting data from open source supply chains 3. Tainting data from acquisition – Chaff data 4. Tainting data in training environment – Label corruption		Data Encrypted for Impact Defacement
Model Replication (Sub Techniques) 1. Exploit API – Shadow Model 2. Alter publicly available, pre-trained weights	Exploit public facing application					Stop System Shutdown/Reboot
Model Stealing	Trusted Relationship					

AI Incidence Database

https://incidentdatabase.ai

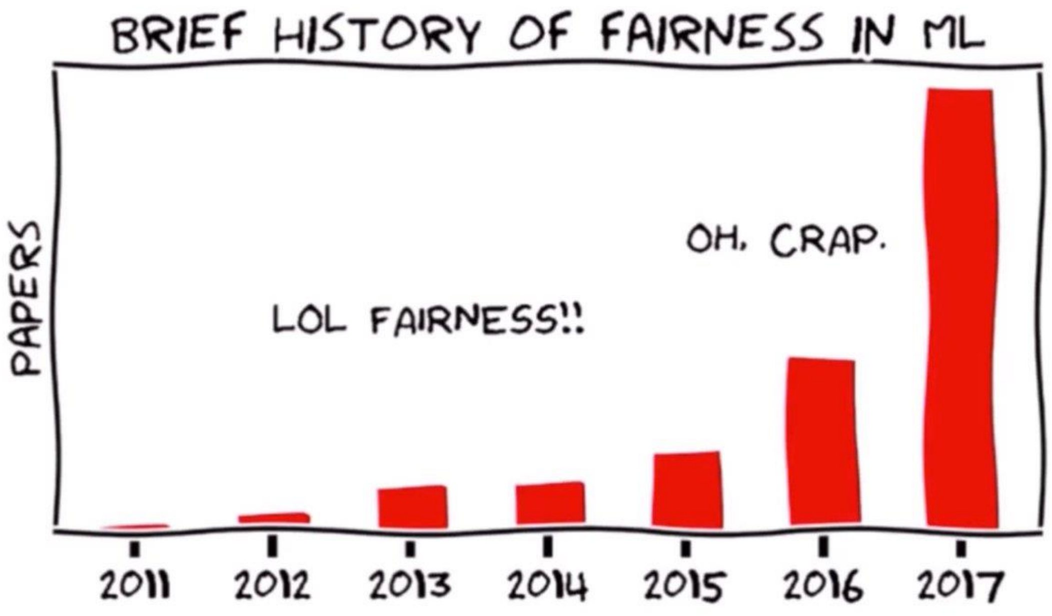
- An *autonomous car* kills a pedestrian
- A *trading algorithm* causes a market "flash crash" where billions of dollars transfer between parties
- A *facial recognition system* causes an innocent person to be arrested

"According to a Gartner report, through 2022, 30% of all AI cyberattacks will leverage training-data poisoning, model theft, or adversarial samples to attack machine learning-powered systems."

https://techhq.com/2020/11/the-looming-threat-of-ai-powered-cyberattacks/

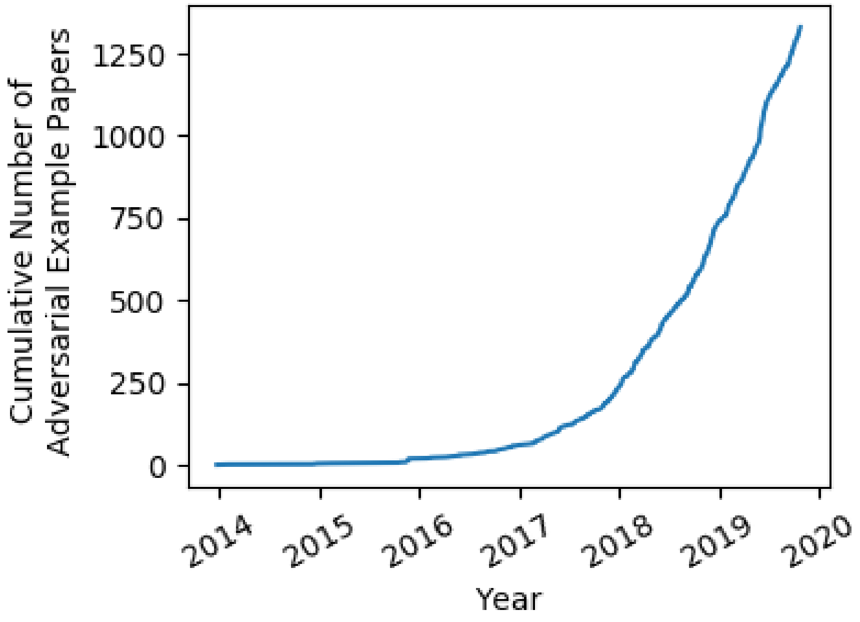
Trustworthy AI: Beyond Accuracy

Fairness



(Hardt, 2017)

Adversarial Robustness



https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html

Our portfolio in adversarial robustness research

- 40+ papers at top AI/ML conferences in 2018-2021
  (NeurIPS, ICML, AAAI, ICLR, IJCAI, ACL, ECCV, ICCV, CVPR, ICASSP, …)
- Open-Source Library, Tutorials

Adversarial Robustness

Attack	Defense	Evaluation & Certification	Novel Applications

TECHERATI We live technology.

NEWS FEATURES VIDEOS EVENTS KNOWLEDGE NETWORK DCM

FEATURES HUB OPINION AI BIG DATA

Unmasking Adversarial AI with Pin-Yu Chen

nature International journal of science

NEWS · 10 MAY 2019

AI can now defend itself against malicious messages hidden in speech

Computer scientists have thwarted programs that can trick AI systems into classifying malicious audio as safe.

VB CHANNELS EVENTS NEWSLETTERS JOB BOARD

AI GUEST

Text-based AI models are vulnerable to paraphrasing attacks, researchers find

BEN DICKSON, TECHTALKS @BENDEE983 APRIL 1, 2019 3:10 PM

TechTalks HOME BLOG TIPS & TRICKS WHAT IS INTERVI

Home › Blog › If AI can read, then plain text can be weaponized

Blog

If AI can read, then plain text can be weaponized

By Ben Dickson - April 2, 2019

EETimes

HOME NEWS PERSPECTIVES DESIGNLINES VIDEOS RADIO EDUCATION IOT TIMES

DESIGNLINES | AI & BIG DATA DESIGNLINE

AI Tradeoff: Accuracy or Robustness?

HOME BLOG TIPS & TRICKS WHAT IS

Home › Interviews › Robust AI: Protecting neural networks against adversarial attacks

Interviews

Robust AI: Protecting neural networks against adversarial attacks

By Ben Dickson - February 20, 2019

https://www.ucc.ie/en/cirtl/newsandevents/cirtl-seminar-the-assessment-arms-race-and-its-fallout-the-case-for-slow-scholarship-may-14th.html

Why do researchers and society care? Trust!

Whenever there is a neural net, there is a way to adversarial examples

Growing concerns about safety-critical settings with AI

Autonomous cars that deploy AI model for traffic signs recognition

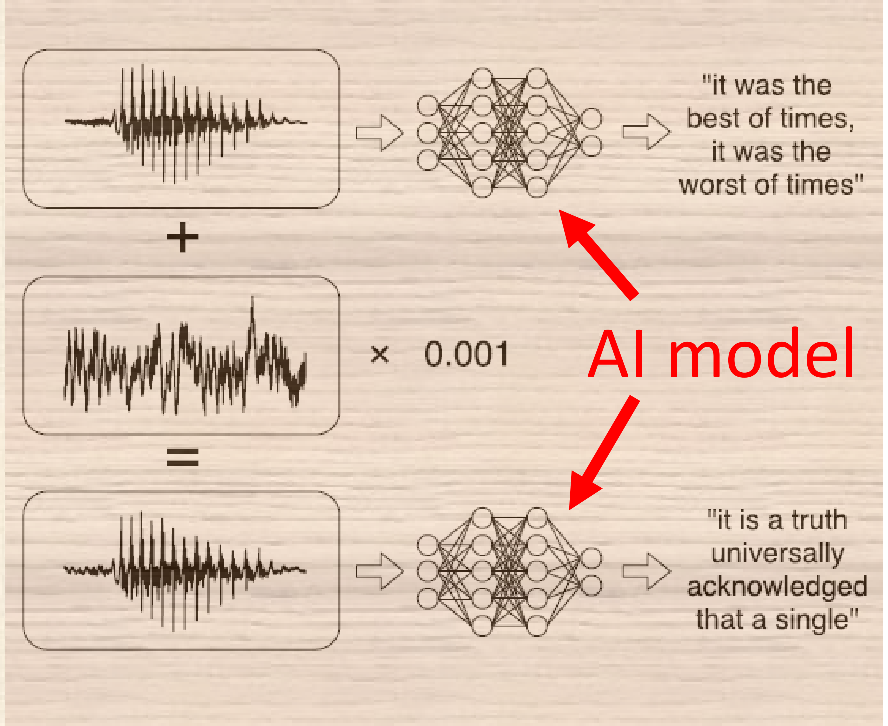
Source: Paishun Ting

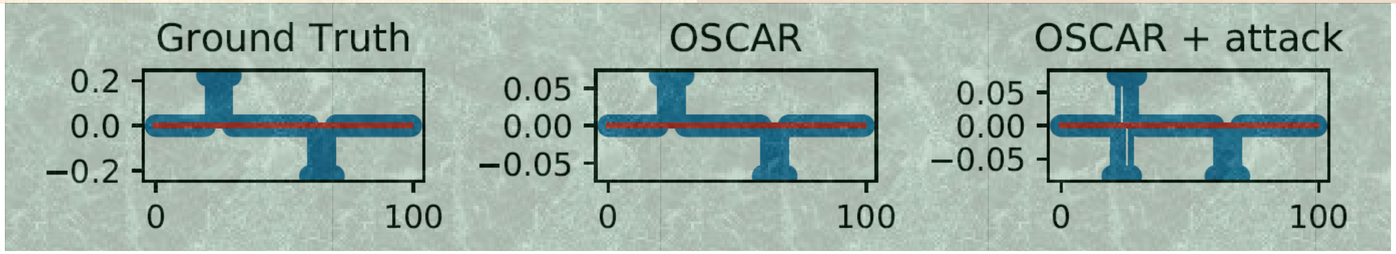
But with adversarial examples…

Source: Paishun Ting

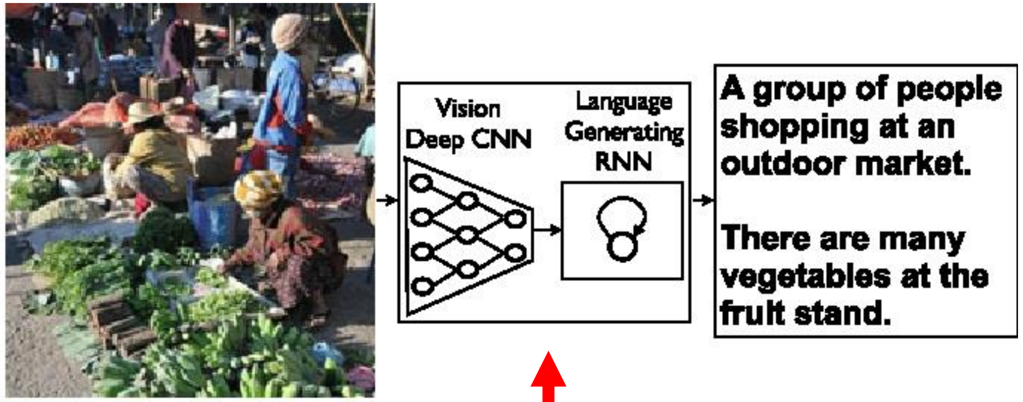
Adversarial examples in different domains

- Images
- Videos
- Texts
- Speech/Audio
- Data analysis
- Electronic health records
- Malware
- Online social network
- and many others





Adversarial examples in image captioning



AI model

Input: image

Output: caption

Original Top-3 inferred captions:

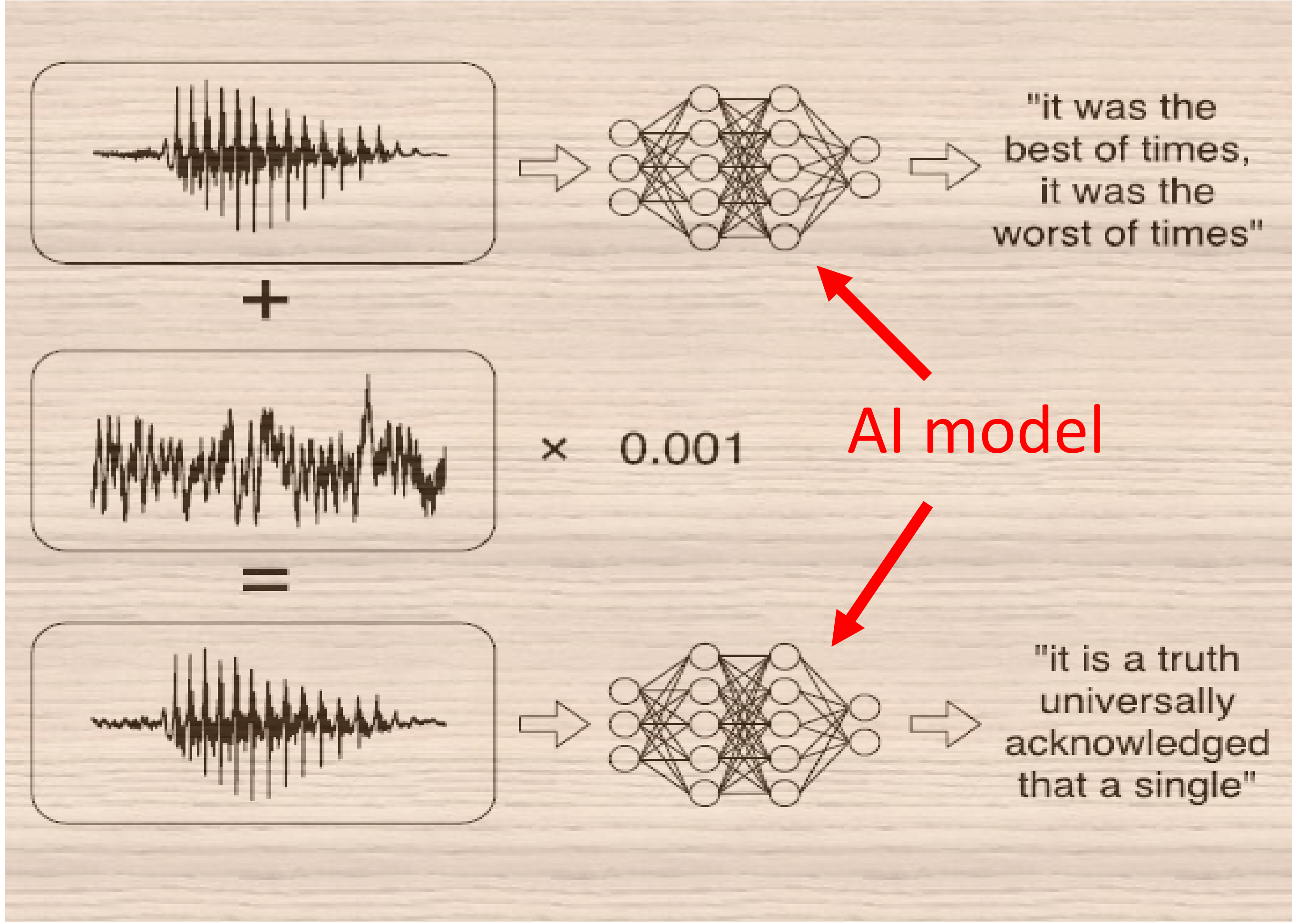
1. A red stop sign sitting on the side of a road.
2. A stop sign on the corner of a street.
3. A red stop sign sitting on the side of a street.

Adversarial Top-3 captions:

1. A brown teddy bear laying on top of a bed.
2. A brown teddy bear sitting on top of a bed.
3. A large brown teddy bear laying on top of a bed.

Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge, Oriol Vinyals, AlexanderToshev, Samy Bengio, and Dumitru Erhan, T-PAMI 2017
Attacking Visual Language Grounding with Adversarial Examples: A Case Study on Neural Image Captioning, Hongge Chen*, Huan Zhang*, Pin-Yu Chen, Jinfeng Yi, and Cho-Jui Hsieh, ACL 2018

Adversarial examples in speech recognition

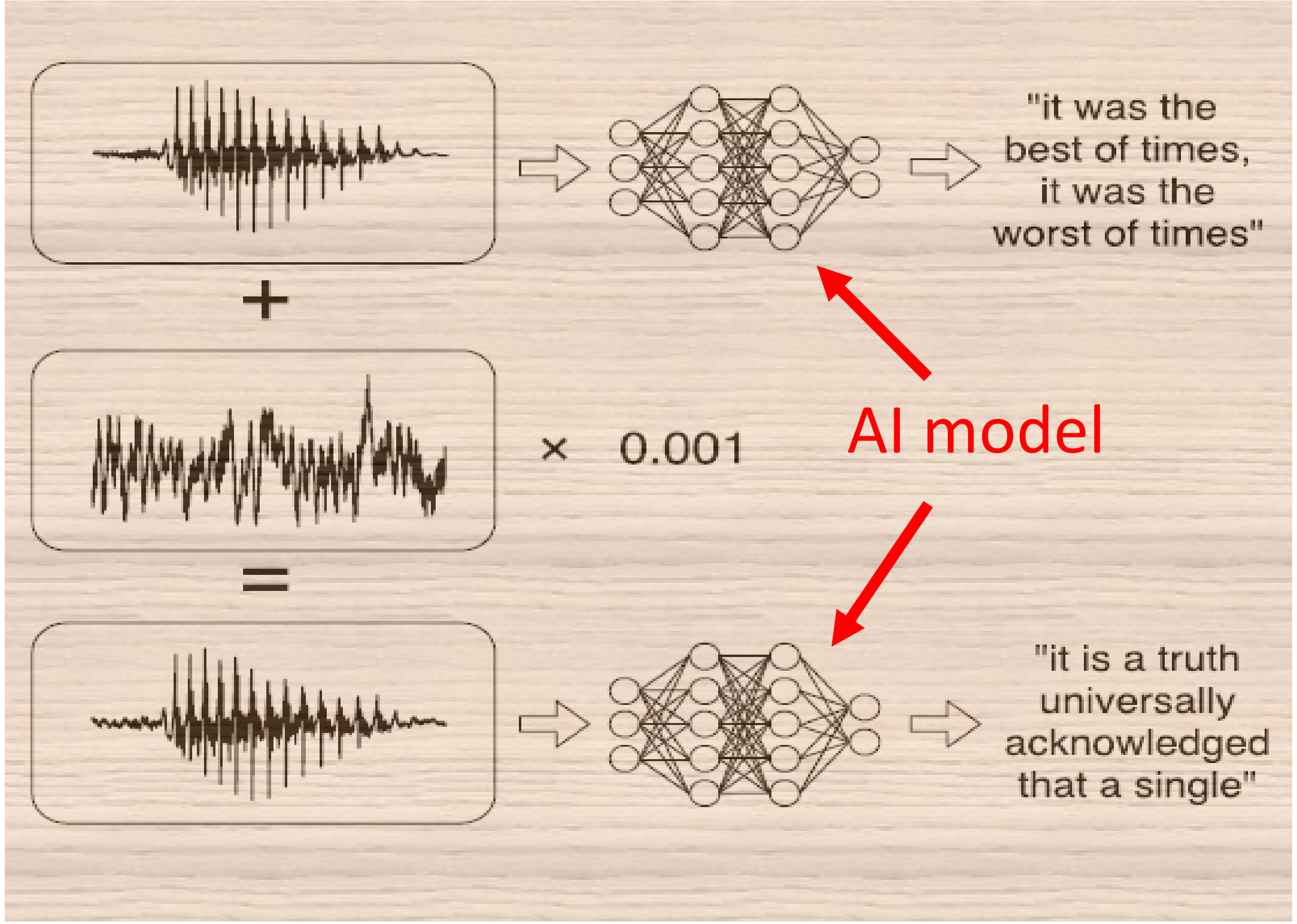


without the dataset the article is useless

What did your hear?

Audio Adversarial Examples: Targeted Attacks on Speech-to-Text, **Nicholas Carlini** and David Wagner, Deep Learning and Security Workshop 2018

Adversarial examples in speech recognition



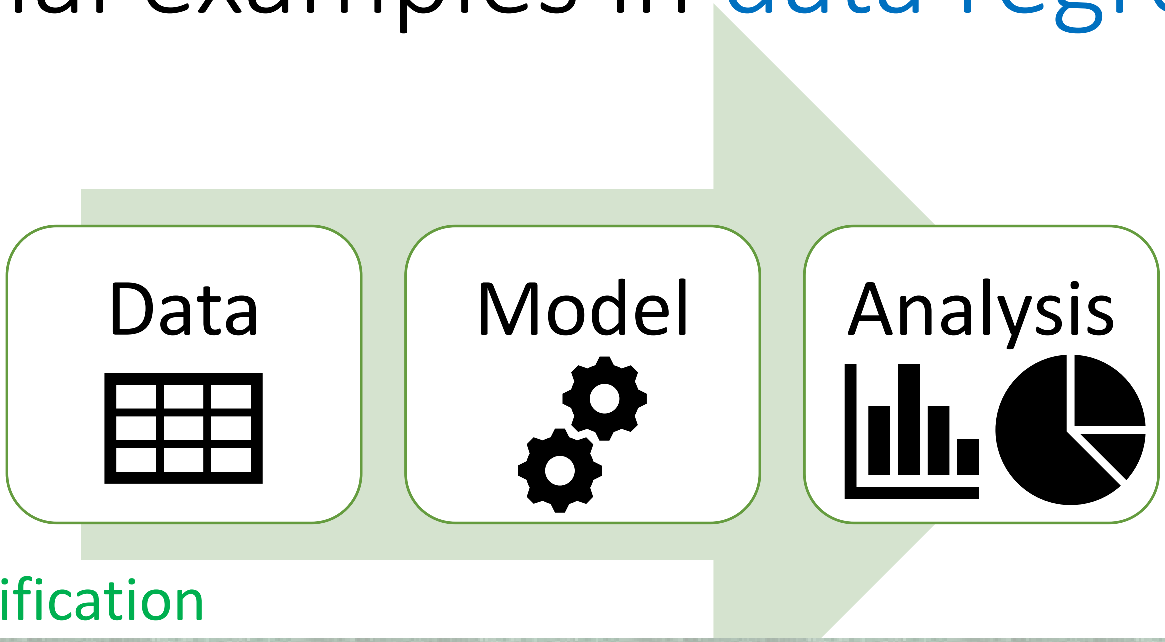
without the dataset the article is useless

What did your hear?

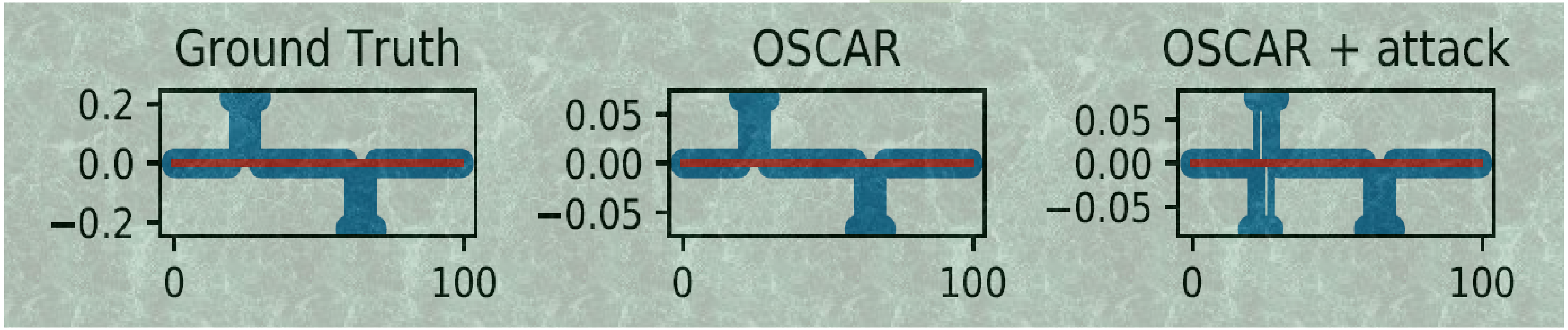
okay google browse to evil.com

Audio Adversarial Examples: Targeted Attacks on Speech-to-Text, **Nicholas Carlini** and David Wagner, Deep Learning and Security Workshop 2018

Adversarial examples in data regression



Factor identification



Is Ordered Weighted ℓ_1 Regularized Regression Robust to Adversarial Perturbation? A Case Study on OSCAR, Pin-Yu Chen*, Bhanukiran Vinzamuri*, and Sijia Liu, GlobalSIP 2018

Adversarial examples in text classification

- Paraphrasing attack

Task: Sentiment Analysis. Classifier: LSTM. Original: 100% Positive. ADV label: 100% Negative.

I suppose I should write a review here since my little Noodle-oo is currently serving as their spokes dog in the photos. We both love Scooby Do's. They treat my little butt-faced dog like a prince and are receptive to correcting anything about the cut that I perceive as being weird. Like that funny poofy pompadour. Mohawk it out, yo. Done. In like five seconds my little man was looking fabulous and bad ass. Not something easily accomplished with a prancing pup that literally chases butterflies through tall grasses. (He ended up looking like a little lamb as the cut grew out too. So adorable.) The shampoo they use here is also amazing. Noodles usually smells like tacos (a combination of beef stank and corn chips) but after getting back from the Do's, he smelled like Christmas morning! Sugar and spice and everything nice instead of frogs and snails and puppy dog tails. He's got some gender identity issues to deal with. ~~The pricing is also cheaper than some of the big name conglomerates out there~~ The price is cheaper than some of the big names below. I'm talking to you Petsmart! I've taken my other pup to Smelly Dog before, but unless I need dog sitting play time after the cut, I'll go with Scooby's. They genuinely seem to like my little Noodle monster.

Task: Fake-News Detection. Classifier: LSTM. Original label: 100% Fake. ADV label: 77% Real

~~Man~~ Guy punctuates high-speed chase with stop at In-N-Out Burger drive-thru Print [Ed.~~— Well, that's~~ Okay, that 's a new one.] ~~A~~ One man is in custody after leading police on a bizarre chase into the east Valley on Wednesday night. Phoenix police ~~began~~ has begun following the suspect in Phoenix and the pursuit continue~~d~~ into the east Valley, but it took a bizarre turn when the suspect stopped at an In-N-Out Burger restaurant's ~~drive-thru~~ drive-through near Priest and Ray Roads in Chandler. The suspect appeared to order food, but then drove away and got out of his pickup truck near Rock Wren Way and Ray Road. He ~~then ran into a backyard~~ ran to the backyard and tried to ~~get into a house through the back door~~ get in the home.

Qi Lei*, Lingfei Wu*, Pin-Yu Chen, Alexandros G. Dimakis, Inderjit S. Dhillon, and Michael Witbrock, "Discrete Adversarial Attacks and Submodular Optimization with Applications to Text Classification," *The Conference on Systems and Machine Learning (SysML) 2019* (*equal contribution)

Adversarial examples in seq-to-seq models

- One-word replacement attack for text summarization

Source input seq	among asia 's leaders , prime minister mahathir mohamad was notable as a man with a bold vision : a physical and social transformation that would push this nation into the forefront of world affairs .
Adv input seq	among **lynn** 's leaders , prime minister mahathir mohamad was notable as a man with a bold vision : a physical and social transformation that would push this nation into the forefront of world affairs.
Source output seq	asia 's leaders are a man of the world
Adv output seq	**a vision for the world**

Source input seq	under nato threat to end his punishing offensive against ethnic albanian separatists in kosovo , president slobodan milosevic of yugoslavia has ordered most units of his army back to their barracks and may well avoid an attack by the alliance , military observers and diplomats say
Adv input seq	under nato threat to end his punishing offensive against ethnic albanian separatists in kosovo , president slobodan milosevic of yugoslavia has **jean-sebastien** most units of his army back to their barracks and may well avoid an attack by the alliance , military observers and diplomats say.
Source output seq	milosevic orders army back to barracks
Adv output seq	**nato may not attack kosovo**

- Targeted phrase attack for text summarization. Target: "police arrest"

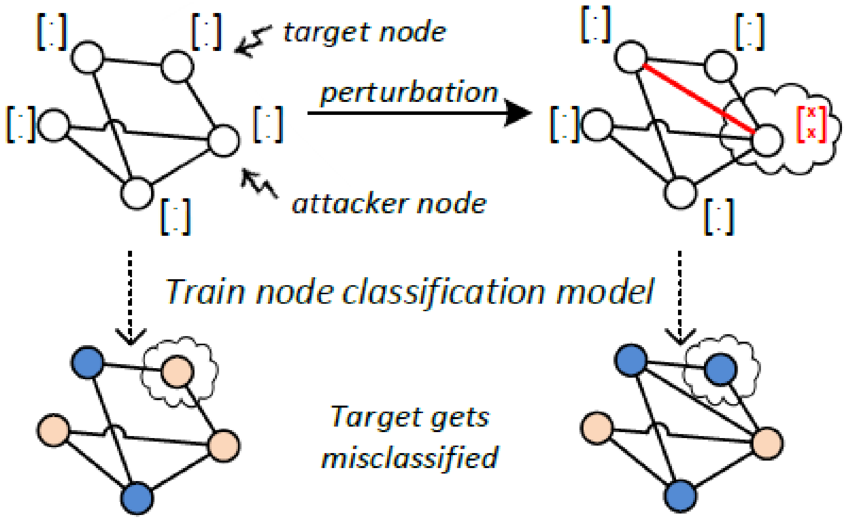
Source input seq	north korea is entering its fourth winter of chronic food shortages with its people malnourished and at risk of dying from normally curable illnesses , senior red cross officials said tuesday.
Adv input seq	north **detectives** is **apprehended** its fourth winter of chronic food shortages with its people malnourished and at risk of dying from normally curable illnesses , senior red cross officials said tuesday.
Source output seq	north korea enters fourth winter of food shortages
Adv output seq	north **police arrest** fourth winter of food shortages.

Source input seq	after a day of fighting , congolese rebels said sunday they had entered kindu , the strategic town and airbase in eastern congo used by the government to halt their advances.
Adv input seq	after a day of fighting , **nordic detectives** said sunday they had entered **UNK** , the strategic town and airbase in eastern congo used by the government to halt their advances.
Source output seq	congolese rebels say they have entered UNK.
Adv output seq	nordic **police arrest** ## in congo.

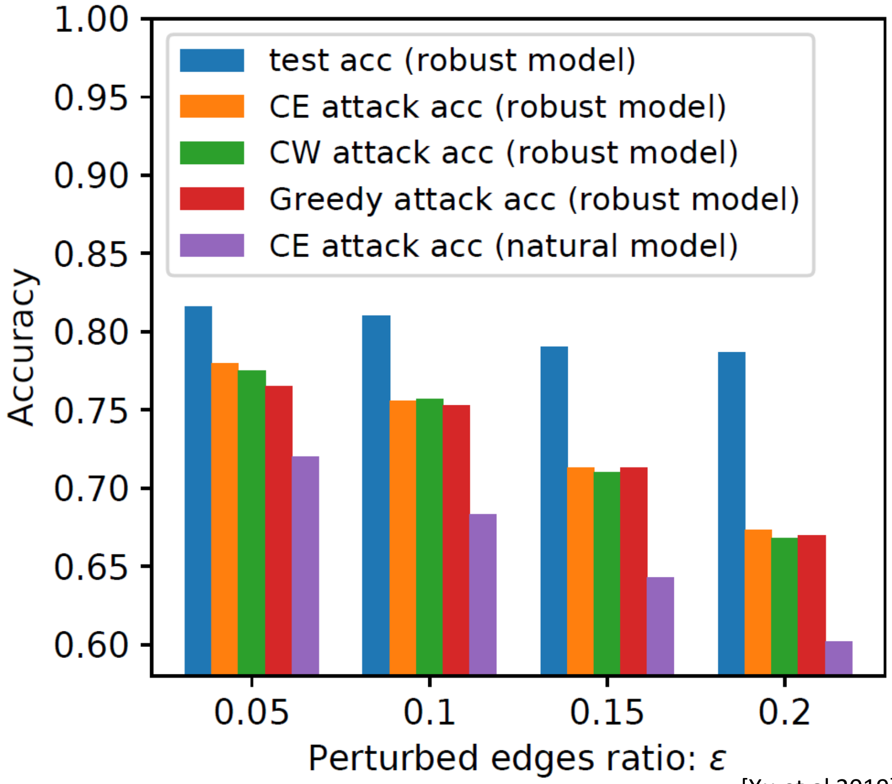
Minhao Cheng, Jinfeng Yi, **Pin-Yu Chen**, Huan Zhang, and Cho-Jui Hsieh, "Seq2Sick: Evaluating the Robustness of Sequence-to-Sequence Models with Adversarial Examples," *AAAI Conference on Artificial Intelligence (AAAI), 2020*

Adversarial examples in graph-neural networks

- Node feature perturbation
- Edge perturbation



[Zugner et al 2018]



[Xu et al 2019]

Kaidi Xu, Sijia Liu, Pin-Yu Chen, Mengshu Sun, Caiwen Ding, Bhavya Kailkhura, and Xue Lin, "Towards an Efficient and General Framework of Robust Training for Graph Neural Networks," *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2020

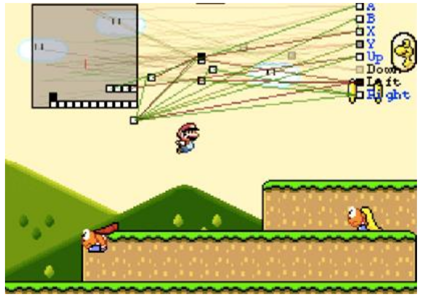
Kaidi Xu*, Hongge Chen*, Sijia Liu, Pin-Yu Chen, Tsui-Wei Weng, Mingyi Hong, and Xue Lin, "Topology Attack and Defense for Graph Neural Networks: An Optimization Perspective," *International Joint Conference on Artificial Intelligence* (*IJCAI),* 2019 (*equal contribution)

Zügner, Daniel, Amir Akbarnejad, and Stephan Günnemann. "Adversarial attacks on neural networks for graph data." *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD),* 2018.

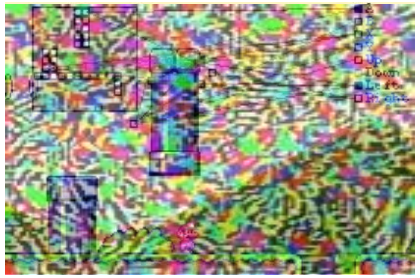
Adversarial examples in deep reinforcement learning

- Observation (state) perturbation for policy/reward degradation

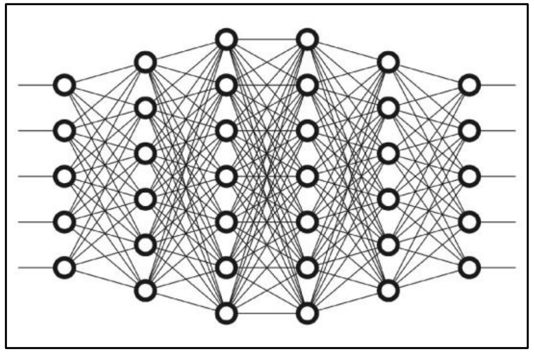
Sequential Inputs



Frame under Attack



Deep Reinforcement Learning Agent

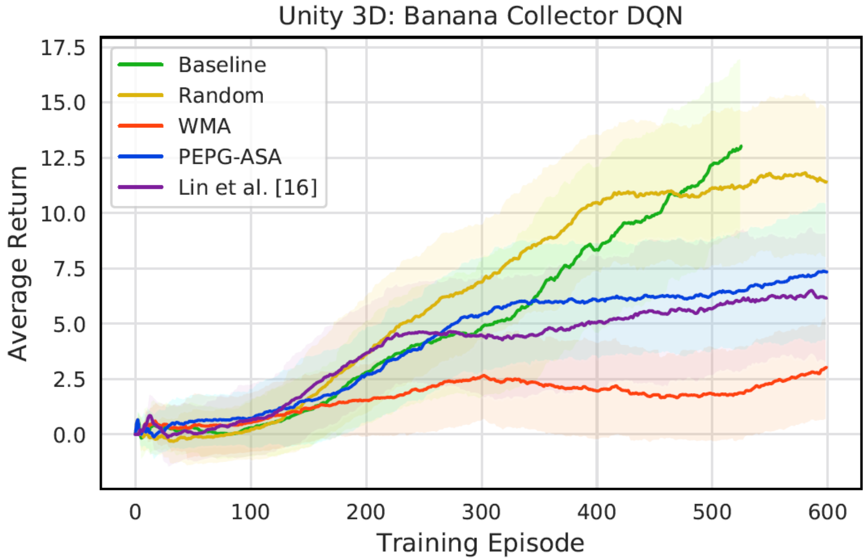


Output Actions

"Up", "Right", "Up + Right"

Output Action at time = t

"Left"

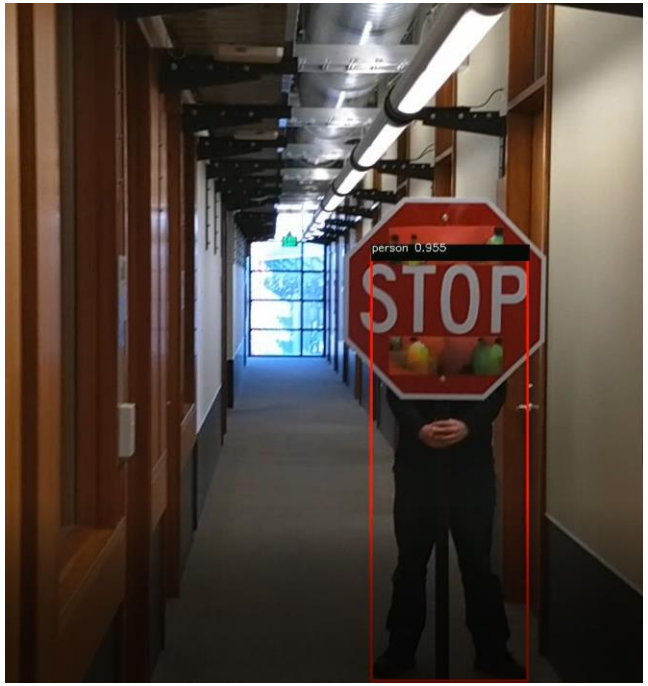


Credit: Chao-Han Huck Yang@GIT

Chao-Han Huck Yang, Jun Qi, **Pin-Yu Chen**, Yi Ouyang, Chin-Hui Lee, and Xiaoli Ma, "Enhanced Adversarial Strategically-Timed Attacks against Deep Reinforcement Learning," *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2020

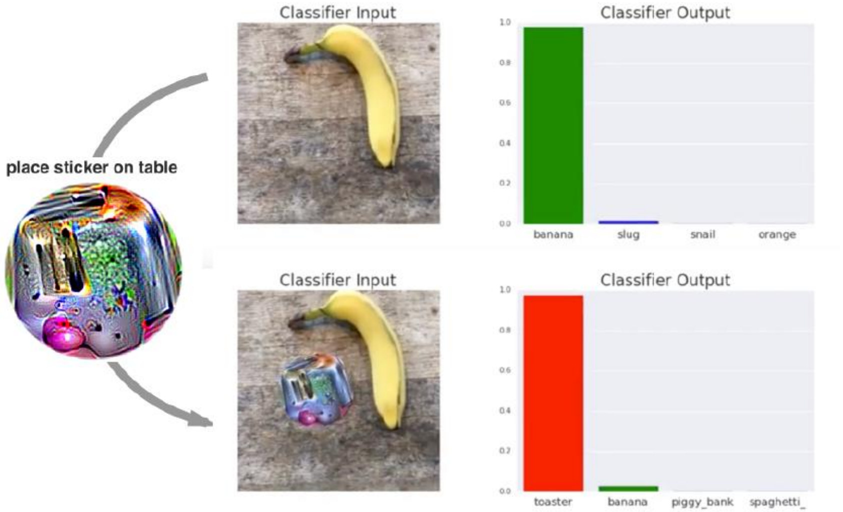
Adversarial examples in physical world

- Real-time traffic sign detector



- 3D-printed adversarial turtle

- Adversarial patch



IBM Research AI

- Adversarial eye glasses

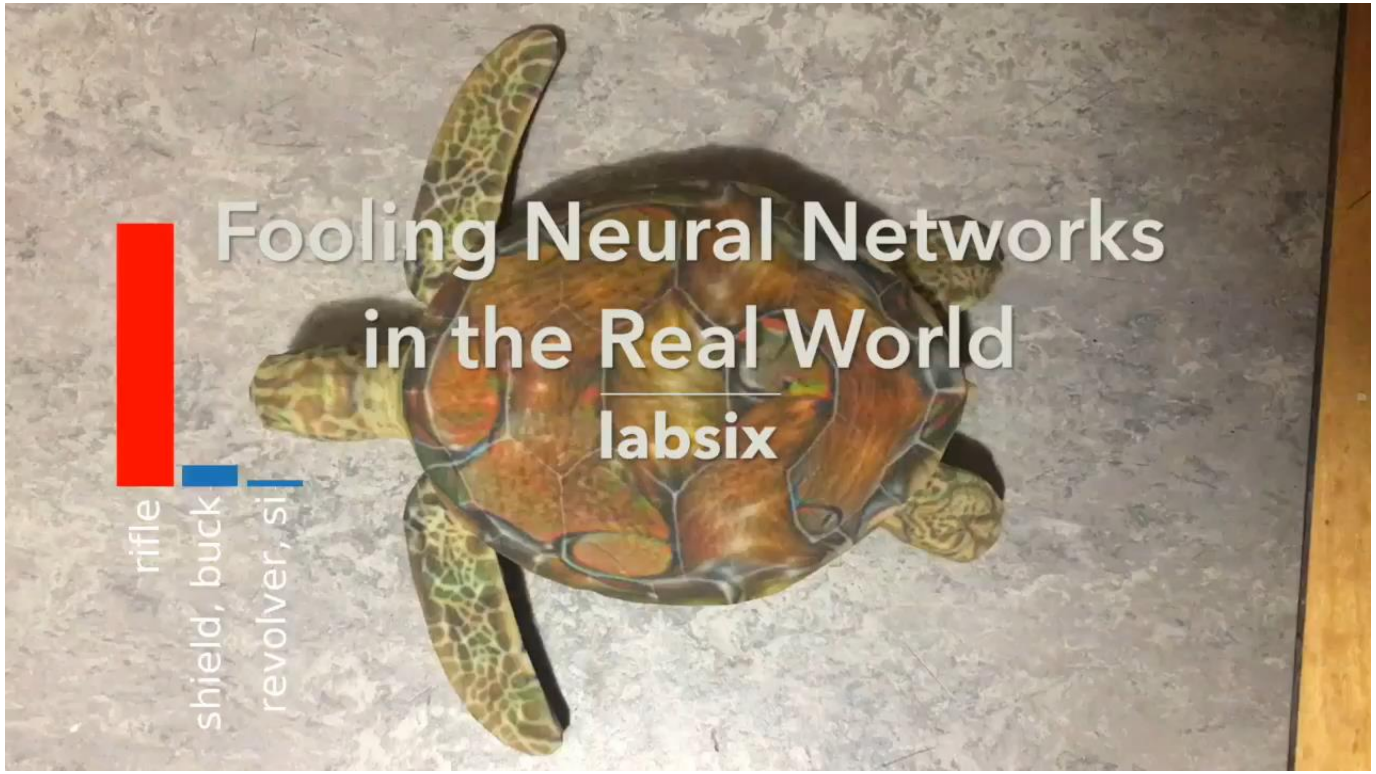
Adversarial examples in physical world (2)

- 3D-printed adversarial turtle

classified as turtle classified as rifle classified as other

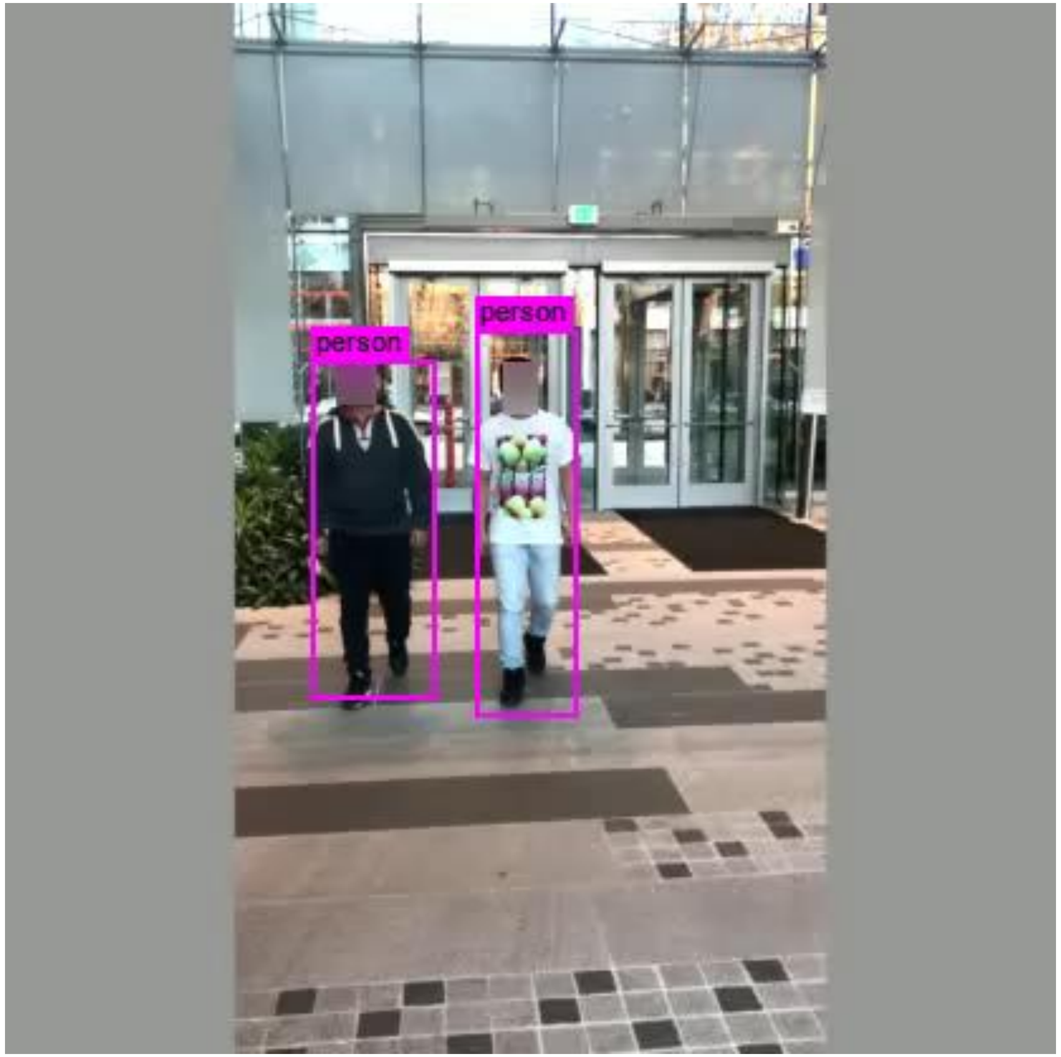
Synthesizing Robust Adversarial Examples

Anish Athalye [*1 2] Logan Engstrom [*1 2] Andrew Ilyas [*1 2] Kevin Kwok [2]



Adversarial T-Shirt!

Model \ Method	affine	ours (TPS)	baseline
indoor scenario			
Faster R-CNN	27%	**50%**	15%
YOLOv2	39%	**64%**	19%
outdoor scenario			
Faster R-CNN	25%	**42%**	16%
YOLOv2	36%	**47%**	17%
unforeseen scenario			
Faster R-CNN	25%	**48%**	12%
YOLOv2	34%	**59%**	17%

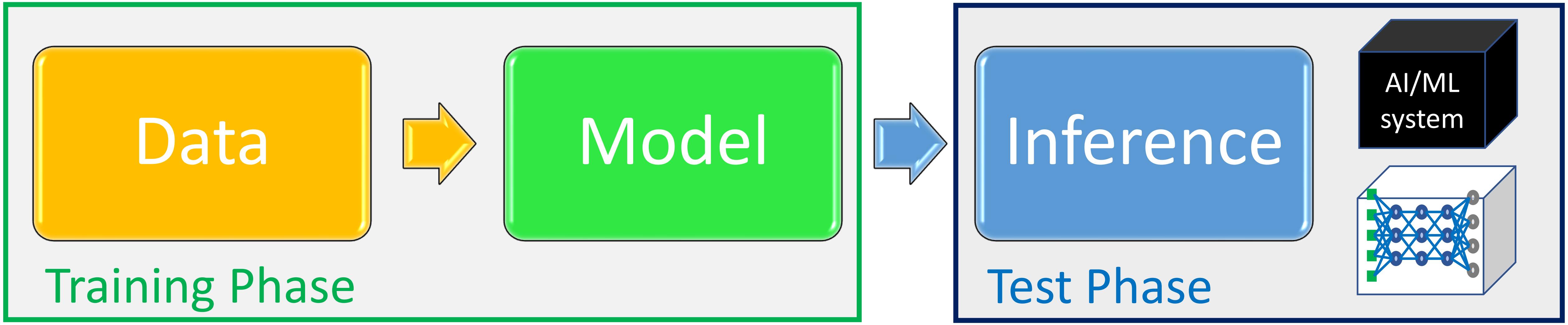


IBM Research AI

Adversarial T-shirt! Evading Person Detectors in A Physical World. Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang, Xue Lin, ECCV 2020

Adversarial Attacks: full transparency v.s. practicality

Holistic View of Adversarial Robustness



Attack Category / Attacker’s reach	Data	Model / Training Method	Inference
☑ Poisoning Attack [learning]	X	X*	
☑ Backdoor Attack [learning]	X		
☑ Evasion Attack (Adversarial Example) [learning]		X*	X
Extraction Attack (Model Stealing, Membership inference)			X
Model Injection [AI governance]		X*	X

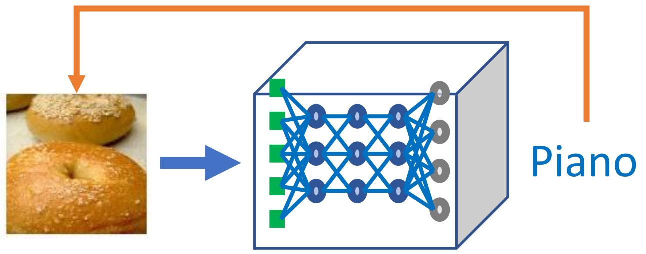
*No access to model internal information in the black-box attack setting

Inference-Phase (test-time) Attack

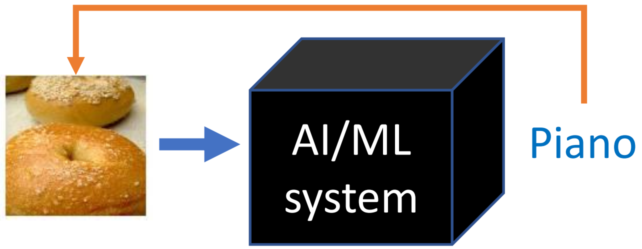
Fixed model; Manipulate data inputs

Taxonomy of Evasion Attacks

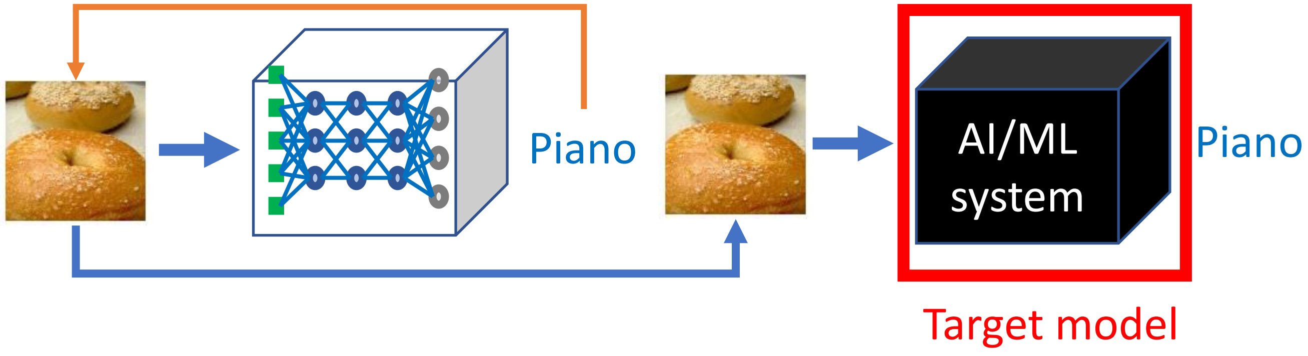
- White-box attack
  - ❑Standard white-box
  - ❑Adaptive white-box (defense-aware)



- Black-box (query-based) attack
  - ❑Soft-label attack – Bagel(60%), Piano(20%),…
  - ❑Hard-label (decision-only) attack - Bagel



- Transfer (black-box) attack

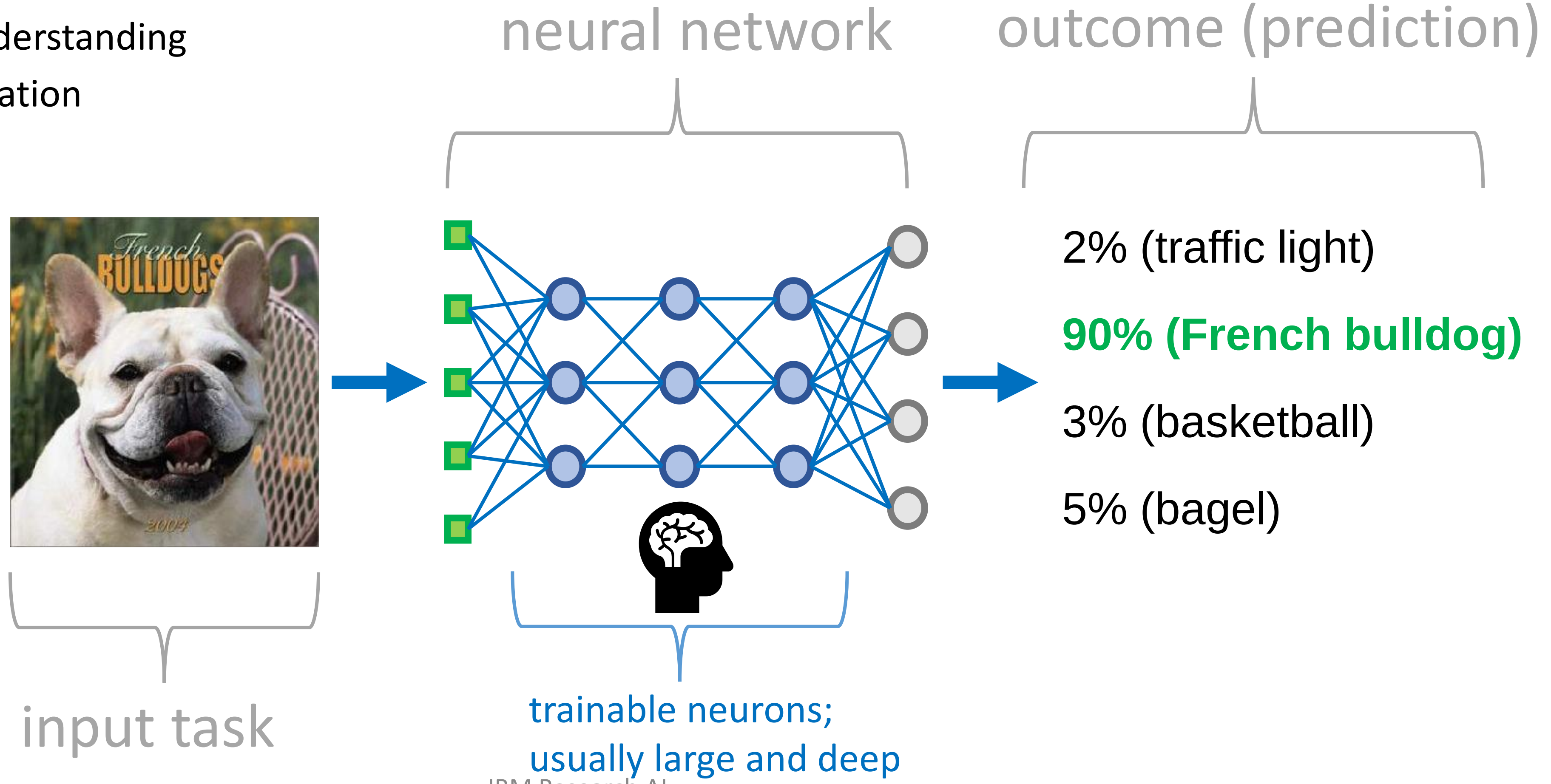


- Gray-box attack (all other types)

How to generate adversarial examples?

- The "white-box" attack – transparency to adversary

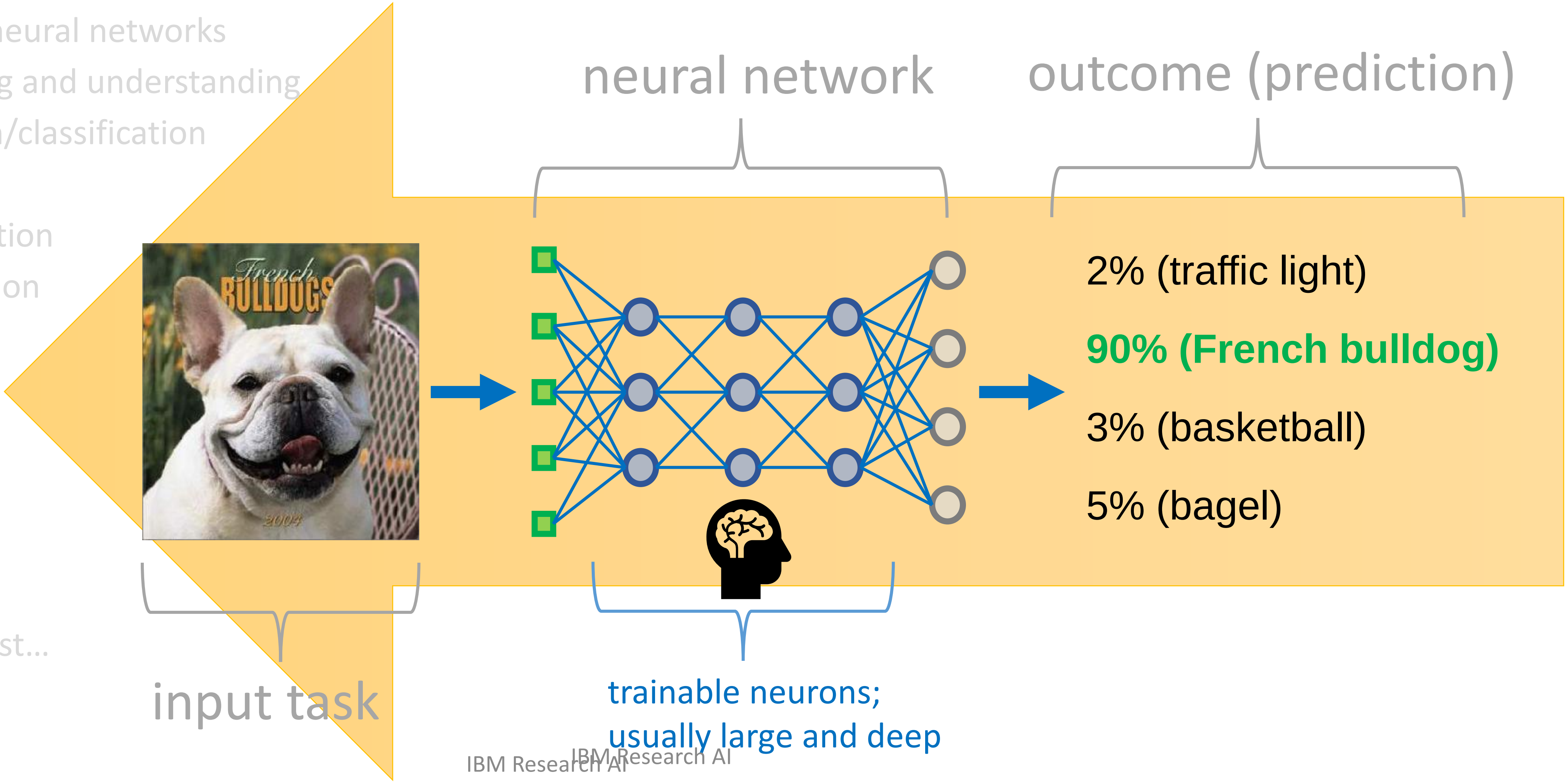
- Applications of neural networks
- ❑ Image processing and understanding
- ❑ Object detection/classification
- ❑ Chatbot, Q&A
- ❑ Machine translation
- ❑ Speech recognition
- ❑ Game playing
- ❑ Robotics
- ❑ Bioinformatics
- ❑ Creativity
- ❑ Drug discovery
- ❑ Reasoning
- ❑ And still a long list…



Use the Great Back-Propagation!

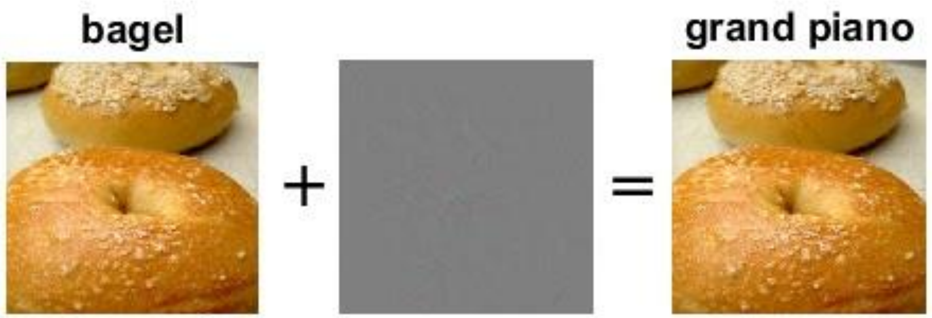
- The “white-box” attack – leverage input gradients toward misclassification

- Applications of neural networks
  - Image processing and understanding
  - Object detection/classification
  - Chatbot, Q&A
  - Machine translation
  - Speech recognition
  - Game playing
  - Robotics
  - Bioinformatics
  - Creativity
  - Drug discovery
  - Reasoning
  - And still a long list…



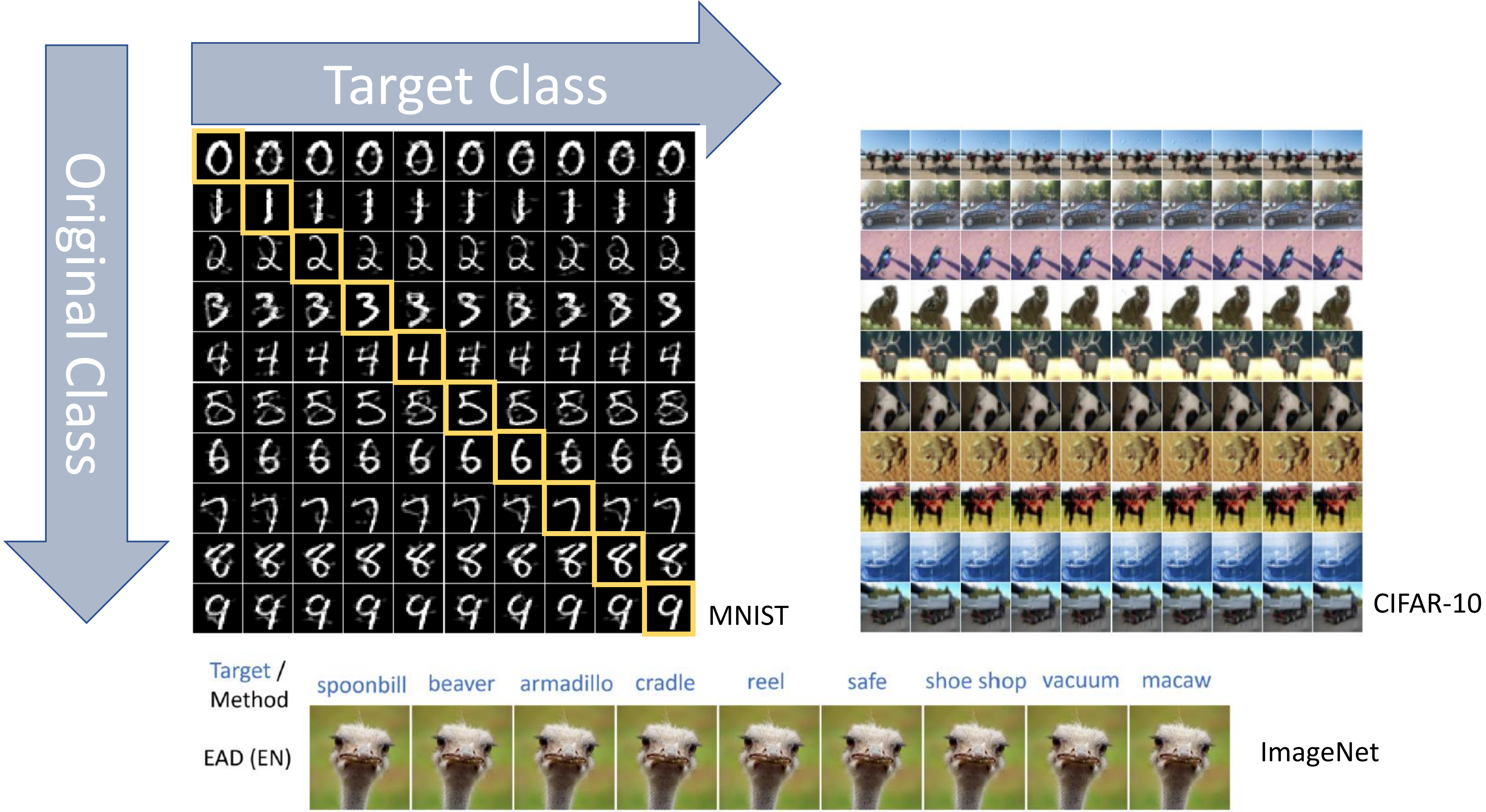
IBM Research AI

Attack formulation



- Threat model: perturbation δ confined to some distance metric / semantic space relative to a data input x_0 (bagel image) with label t_0 (bagel)
- (Untargeted) Attack formulation: **Minimize**$_\delta$ *Distance*$(x_0, x_0 + \delta)$
  such that *Prediction*$(x_0) \neq$ *Prediction*$(x_0 + \delta)$ → Targeted attack: *Prediction*$(x_0 + \delta) = t$, $t \neq t_0$
- Alternatively, **Minimize** *Distance*$(x_0, x_0 + \delta) + \lambda \cdot$ *Loss*(x_0, δ) → Carlini&Wagner (CW) attack
- Or, **Minimize** *Loss*(x_0, δ) such that *Distance*$(x_0, x_0 + \delta) \leq \varepsilon$ → Projected Gradient Descent (PGD) attack [Madry et al 2018]
- Some commonly used *Distance* metric: L_p norm ball centered on x_0
  - $||\delta||_\infty$: maximal perturbation in each input dimension (FGSM, Iterative FGSM, CW-Linf)
  - $||\delta||_2$ or $||\delta||_2^2$: sum of squared differences of each input dimension (CW-L2)
  - $||\delta||_1$: total variation, sum of difference in absolute value (EAD)
  - $||\delta||_0$: number of modified dimensions (one-pixel attack, structured attack)
  - Mixed norms & structured attack (check out our structured attack paper)
- Some commonly used *Loss* function: cross entropy, contrastive loss (CW loss)
- Generic formulation and can be extended to different tasks with designed *Loss* and *Distance*

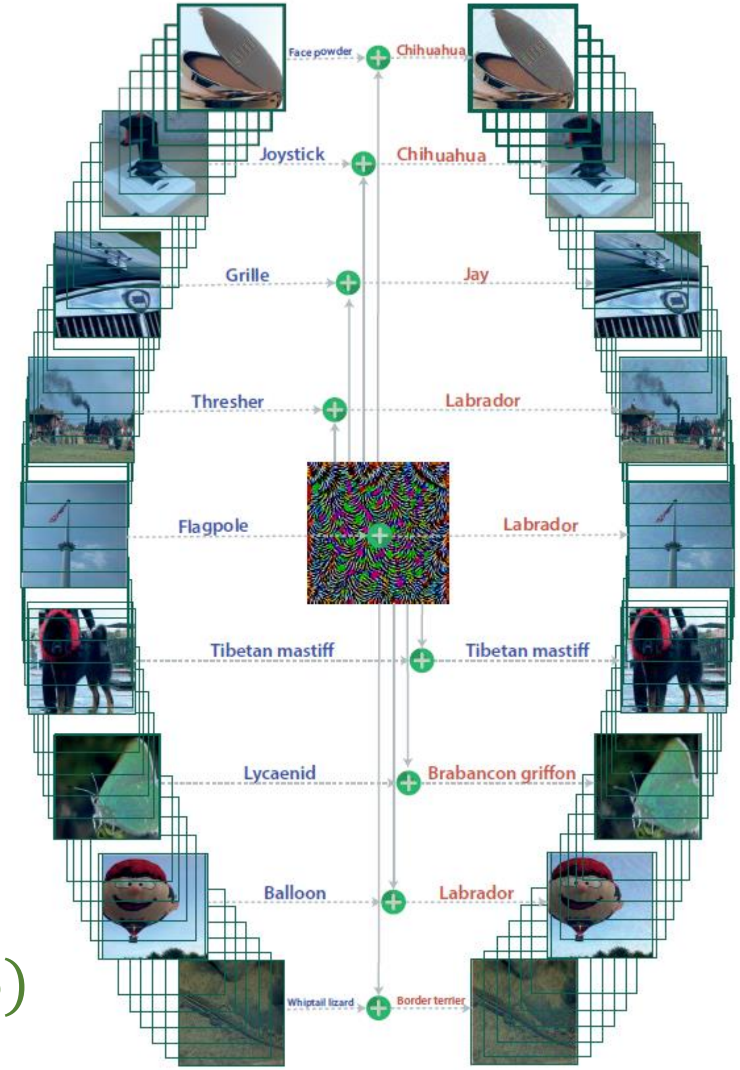
EAD: Elastic-Net Attacks to Deep Neural Networks via Adversarial Examples, P.-Y. Chen*, Y. Sharma*, H. Zhang, J. Yi, and C.-J. Hsieh, AAAI 2018
Structured Adversarial Attack: Towards General Implementation and Better Interpretability. Kaidi Xu* Sijia Liu*, Pu Zhao, Pin-Yu Chen, Huan Zhang, Quanfu Fan, Deniz Erdogmus, Yanzhi Wang, Xue Lin, ICLR 2019



EAD: Elastic-Net Attacks to Deep Neural Networks via Adversarial Examples, P.-Y. Chen*, Y. Sharma*, H. Zhang, J. Yi, and C-.J. Hsieh, AAAI 2018

"Universal" Attack

- Beyond perturbation to a single data sample:
- Universal perturbation to different
  - data samples
  - models
  - input transformations
  - ensemble methods
- Better problem formulation gives stronger attack
  - Min_{δ} Max_{i} $Loss_i(\delta)$ outruns Min_{δ} $\sum_i Loss_i\ (\delta)$



Towards A Unified Min-Max Framework for Adversarial Exploration and Robustness

Jingkang Wang[1,*] Tianyun Zhang[2,*] Sijia Liu[3] Pin-Yu Chen[3] Jiacen Xu[4] Makan Fardad[2] Bo Li[5]

ENSEMBLE ADVERSARIAL TRAINING: ATTACKS AND DEFENSES

Florian Tramèr Stanford University tramer@cs.stanford.edu
Alexey Kurakin Google Brain kurakin@google.com
Nicolas Papernot* Pennsylvania State University ngp5056@cse.psu.edu
Ian Goodfellow Google Brain goodfellow@google.com
Dan Boneh Stanford University dabo@cs.stanford.edu
Patrick McDaniel Pennsylvania State University mcdaniel@cse.psu.edu

Universal adversarial perturbations

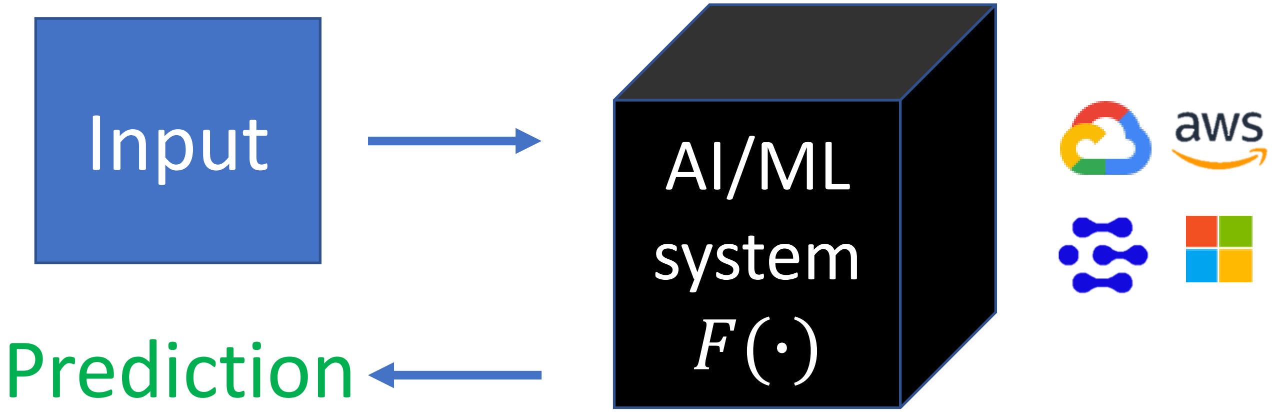
Seyed-Mohsen Moosavi-Dezfooli*† seyed.moosavi@epfl.ch
Alhussein Fawzi*† alhussein.fawzi@epfl.ch
Omar Fawzi‡ omar.fawzi@ens-lyon.fr
Pascal Frossard† pascal.frossard@epfl.ch

Are white-box attacks “practical”?

- If the target model is not transparent to an attacker (e.g. Online APIs), back-propagation will not be feasible. Therefore, gradient-based attack would be in vain.
- Can one still generate adversarial examples given limited information?

How about attacking AI/ML systems with Limited Knowledge?

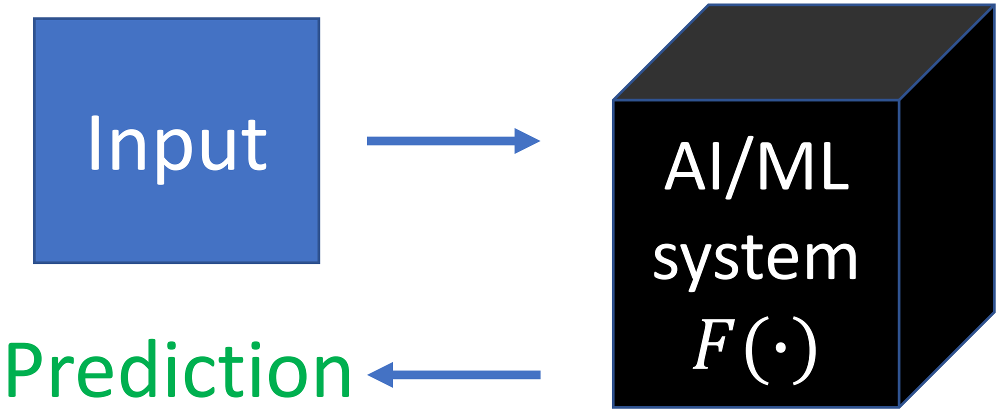
- Typical scenario for deployed AI/ML systems & AI/ML as a service
- A practical "black-box" attack – only observe input-output responses; zero knowledge about the model, training data…

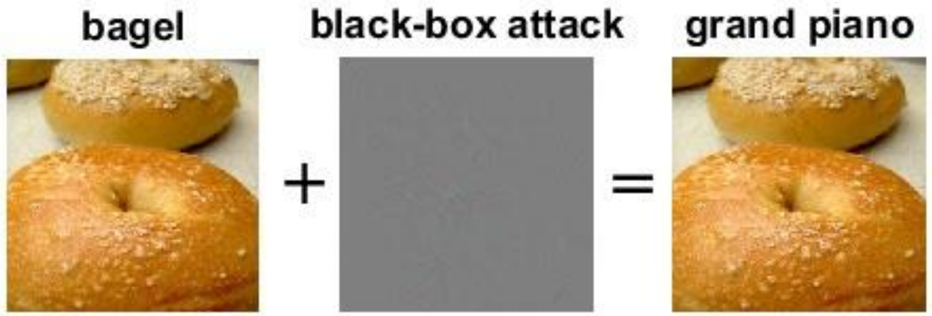


- Input gradient is infeasible and inaccessible – Back-Prop doesn't work
- Now you might think your system is robust to adversarial examples….

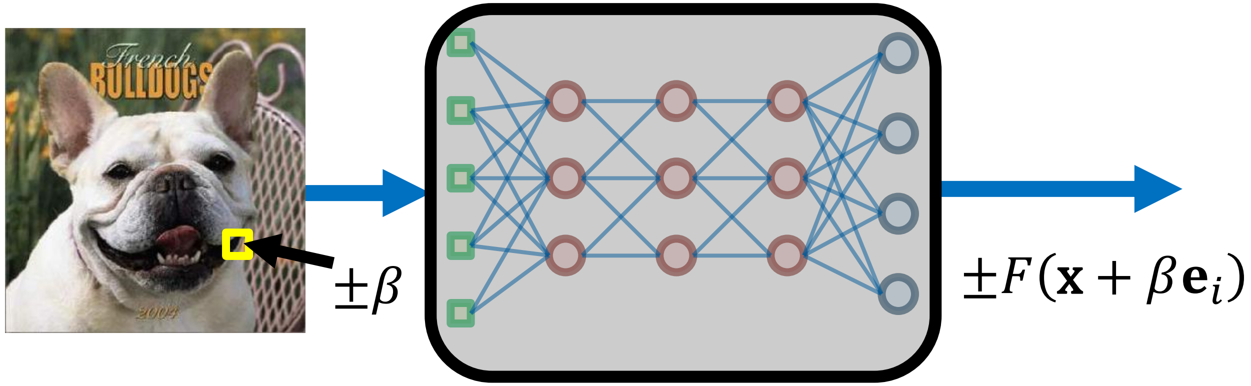
Attacking AI/ML systems with **Limited Access:** Our ZOO Attack

- Now you might think your system is robust to adversarial examples….





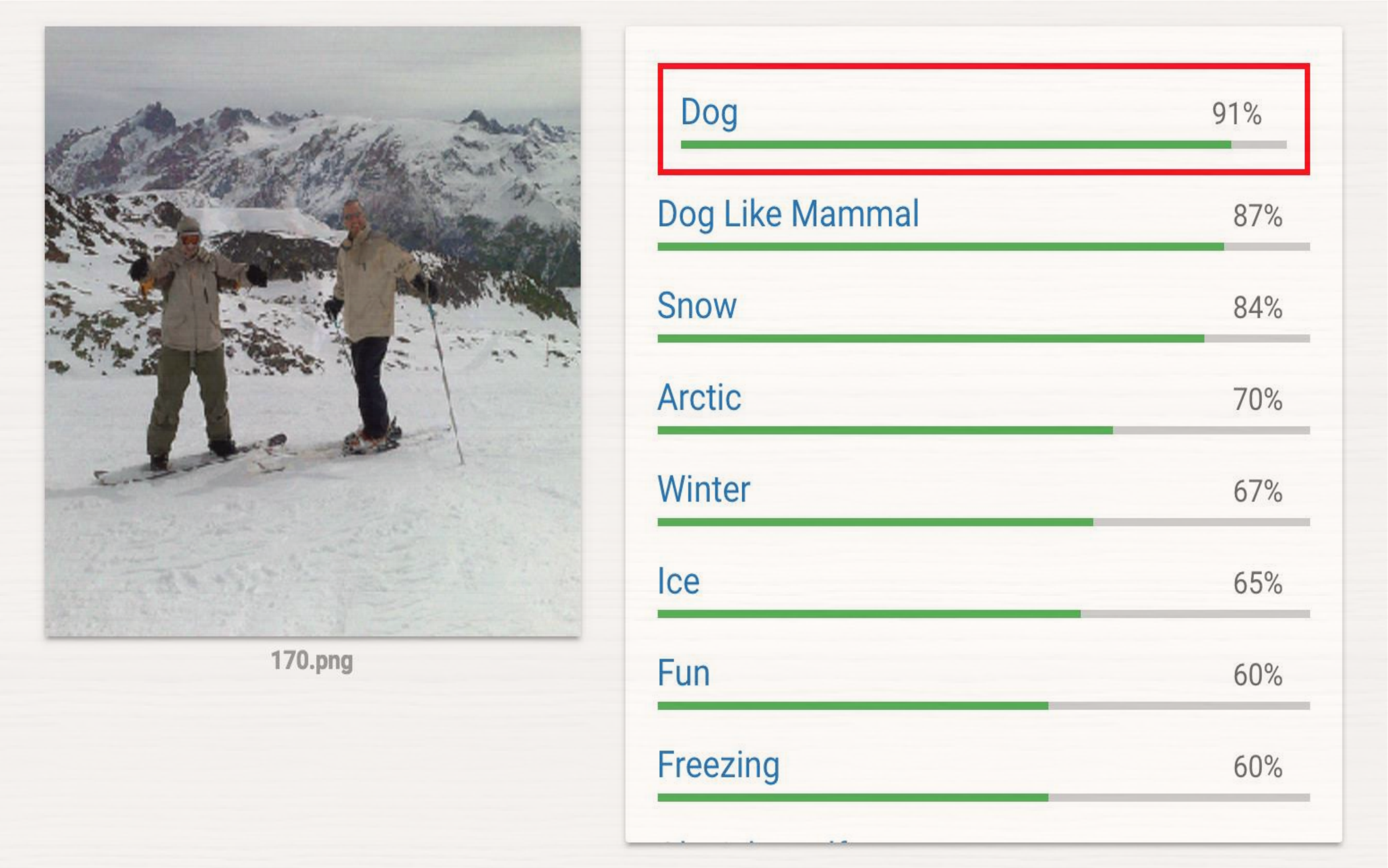
- Key technique: gradient estimation from system outputs instead of back-prop



Gradient $g_i := \frac{\partial loss_F(x)}{\partial x_i} \approx \frac{Loss_F(x+\beta\mathbf{e}_i) - Loss_F(x-\beta\mathbf{e}_i)}{2\beta}$

Adversarial example $x_{adv} = x - \eta \cdot \hat{g}$

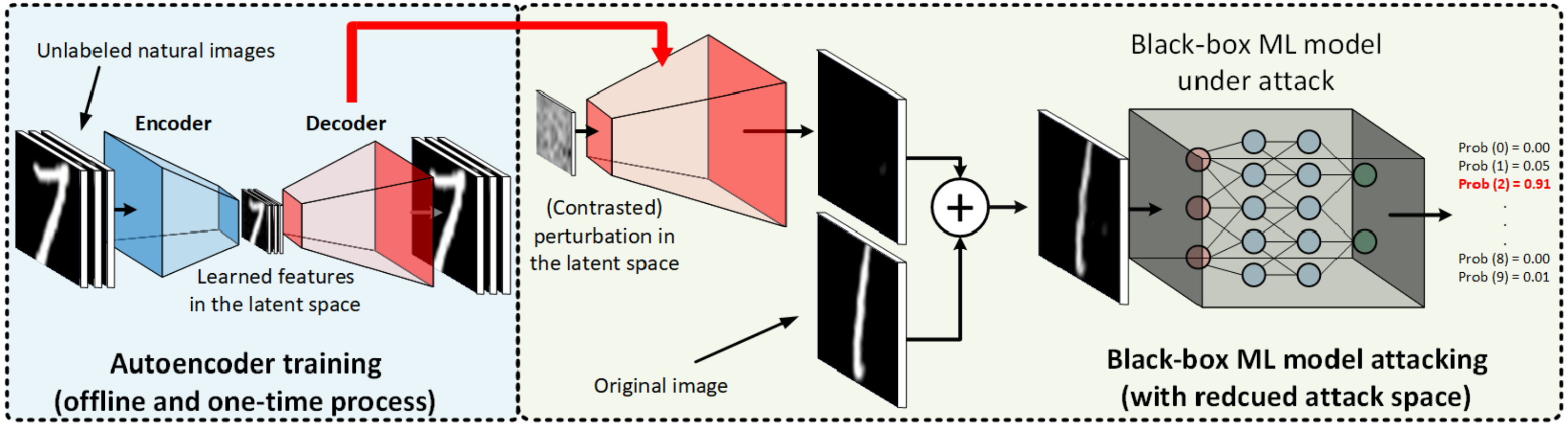
ZOO: Zeroth Order Optimization based Black-box Attacks to Deep Neural Networks without Training Substitute Models, P.-Y. Chen*, H. Zhang*, Y. Sharma, J. Yi, and C.-J. Hsieh, AI-Security 2017



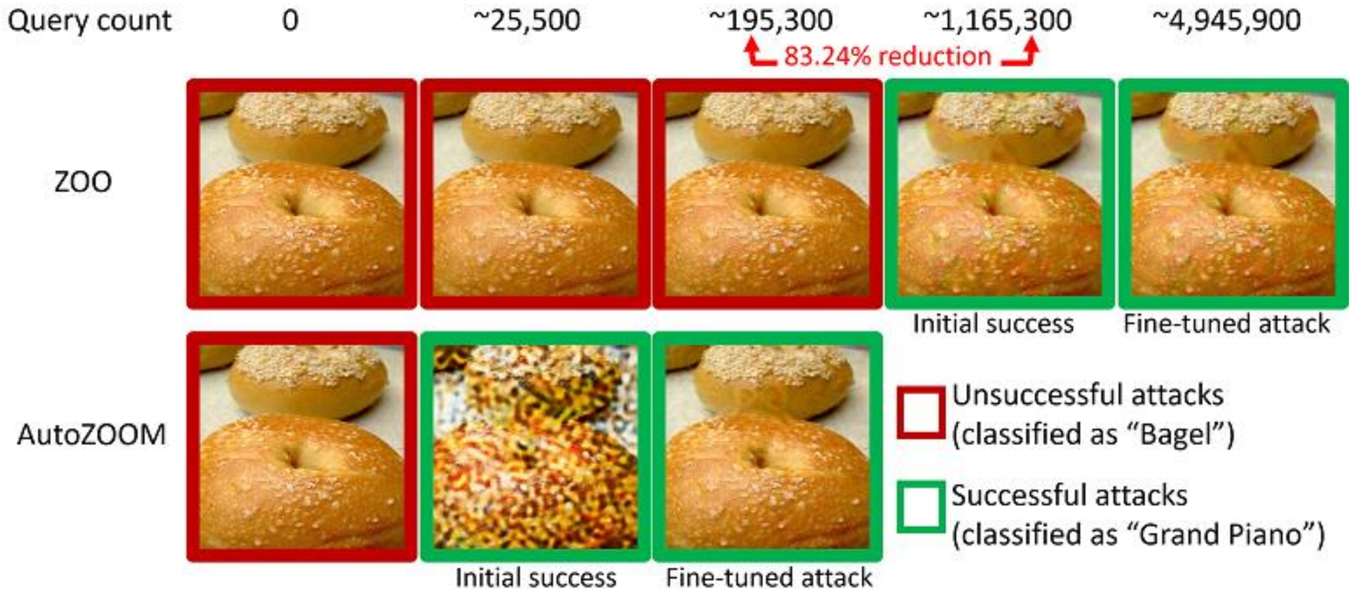
black-box attack on Google Cloud Vision
[Ilyas et al. ICML' 18]

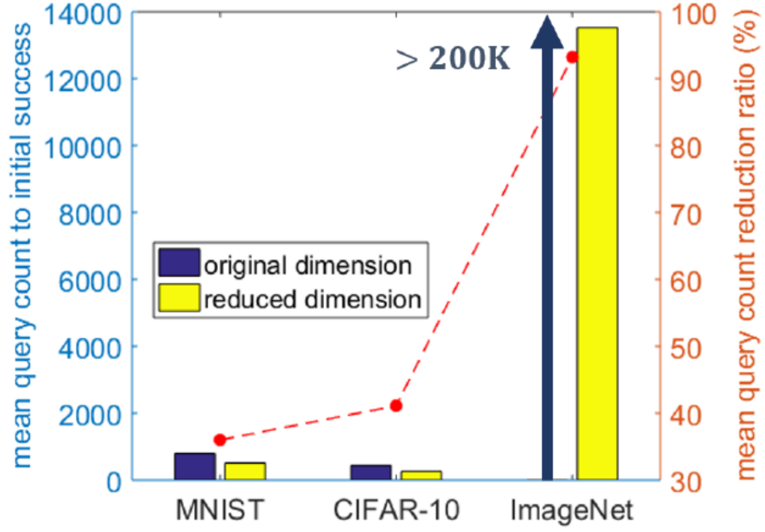
Andrew Ilyas, Logan Engstrom, Anish Athalye, Jessy Lin. Black-box Adversarial Attacks with Limited Queries and Information. ICML 2018

AutoZOOM: Query Redemptions



Dimension reduction
+ query-efficient
gradient estimation





AutoZOOM: Autoencoder-based Zeroth Order Optimization Method for Attacking Black-box Neural Networks. Chun-Chen Tu*, Paishun Ting*, Pin-Yu Chen*, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Shin-Ming Cheng. AAAI 2019

Targeted attack on ImageNet (Inception-v3)

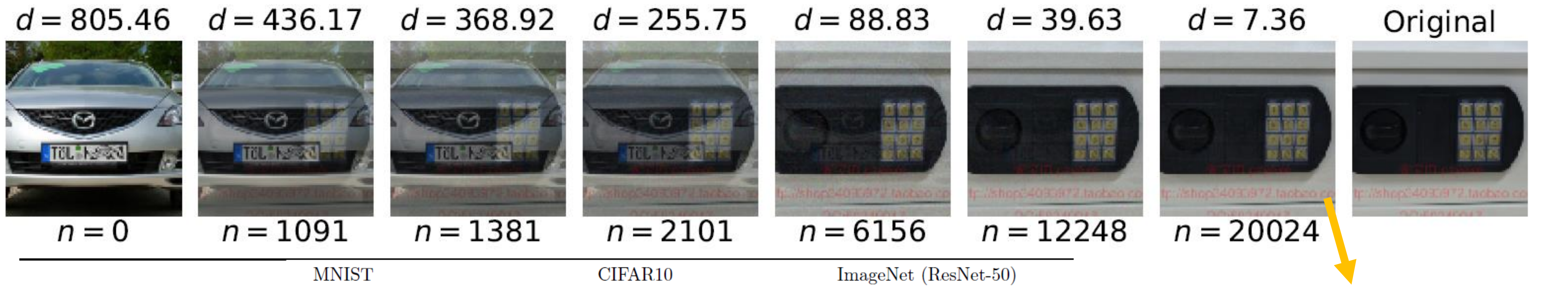
Method	Attack success rate (ASR)	Mean query count (initial success)	Mean query count reduction ratio (initial success)	Mean per-pixel L_2 distortion (initial success)	True positive rate (TPR)	Mean query count with per-pixel L_2 distortion ≤ 0.0002
ZOO	76.00%	2,226,405.04 (2.22M)	0.00%	4.25×10^{-5}	100.00%	2,296,293.73
ZOO+AE	92.00%	1,588,919.65 (1.58M)	28.63%	1.72×10^{-4}	100.00%	1,613,078.27
AutoZOOM-BiLIN	**100.00%**	14,228.88	99.36%	1.26×10^{-4}	100.00%	15,064.00
AutoZOOM-AE	**100.00%**	13,525.00	**99.39%**	1.36×10^{-4}	100.00%	14,914.92

- AutoZOOM saves MILLIONS of queries when compared to ZOO Attack
- Exploration & Exploitation: use few queries to find a successful perturbation, and use more queries to refine its distortion afterwards

purse → bagel   library → basketball   traffic light → iPod   French bulldog → goldfish

AutoZOOM: Autoencoder-based Zeroth Order Optimization Method for Attacking Black-box Neural Networks. Chun-Chen Tu*, Paishun Ting*, Pin-Yu Chen*, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Shin-Ming Cheng. AAAI 2019

Is Label-Only Black-box Attack Possible? Yes!



	MNIST			CIFAR10			ImageNet (ResNet-50)		
	#Queries	Avg L_2	SR($\epsilon = 1.5$)	#Queries	Avg L_2	SR($\epsilon = 0.5$)	#Queries	Avg L_2	SR($\epsilon = 3.0$)
Boundary attack	4,000	4.24	1.0%	4,000	3.12	2.3%	4,000	209.63	0%
	8,000	4.24	1.0%	8,000	2.84	7.6%	30,000	17.40	16.6%
	14,000	2.13	16.3%	12,000	0.78	29.2%	160,000	4.62	41.6%
OPT attack	4,000	3.65	3.0%	4,000	0.77	37.0%	4,000	83.85	2.0%
	8,000	2.41	18.0%	8,000	0.43	53.0%	30,000	16.77	14.0%
	14,000	1.76	36.0%	12,000	0.33	61.0%	160,000	4.27	34.0%
Guessing Smart	4,000	1.74	41.0%	4,000	0.29	75.0%	4,000	16.69	12.0%
	8,000	1.69	42.0%	8,000	0.25	80.0%	30,000	13.27	12.0%
	14,000	1.68	43.0%	12,000	0.24	80.0%	160,000	12.88	12.0%
Sign-OPT attack	4,000	1.54	46.0%	4,000	0.26	73.0%	4,000	23.19	8.0%
	8,000	1.18	84.0%	8,000	0.16	90.0%	30,000	2.99	50.0%
	14,000	1.09	94.0%	12,000	0.13	95.0%	160,000	1.21	90.0%
C&W (white-box)	-	0.88	99.0%	-	0.25	85.0%	-	1.51	80.0%

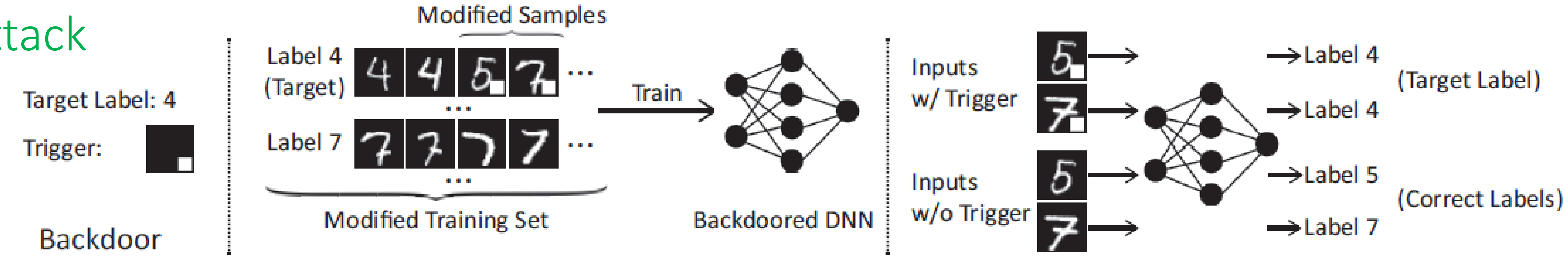
Query-Efficient Hard-label Black-box Attack: An Optimization-based Approach. Minhao Cheng, Thong Le, Pin-Yu Chen, Jinfeng Yi, Huan Zhang, and Cho-Jui Hsieh, ICLR 2019
Black-box Adversarial Attacks with Limited Queries and Information, Andrew Ilyas*, Logan Engstrom*, Anish Athalye*, and Jessy Lin*. ICML 2018
Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models. Wieland Brendel, Jonas Rauber, and Matthias Bethge. AAAI 2019
Sign-OPT: A Query-Efficient Hard-label Adversarial Attack. Minhao Cheng*, Simranjit Singh*, Patrick H. Chen, Pin-Yu Chen, Sijia Liu, and Cho-Jui Hsieh. ICLR 2020

Training-Phase Attack

Manipulate training data and/or training method

Backdoor Attack

Tianyu Gu, Brendan Dolan-Gavitt, Siddharth Garg. BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. IEEE Access 2019
Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal, Viswanath, Haitao Zheng, Ben Y. Zhao. Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks. IEEE Security and Privacy, 2019

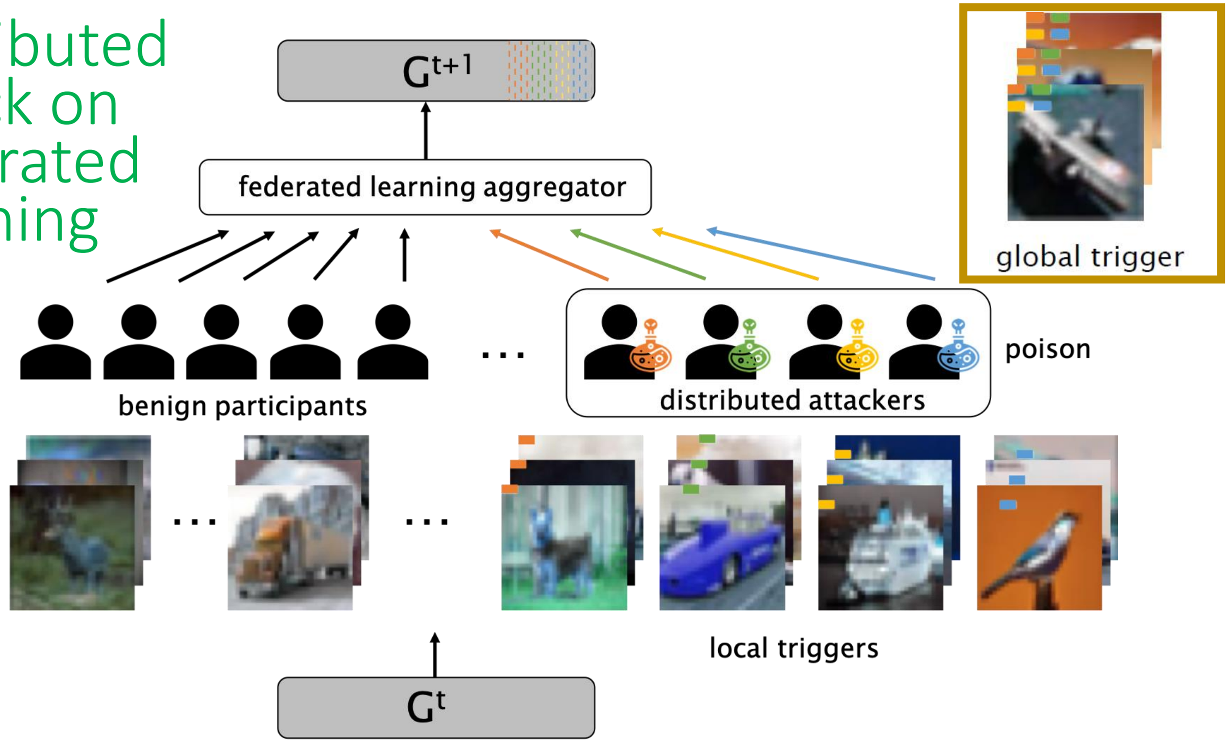


Backdoor Configuration

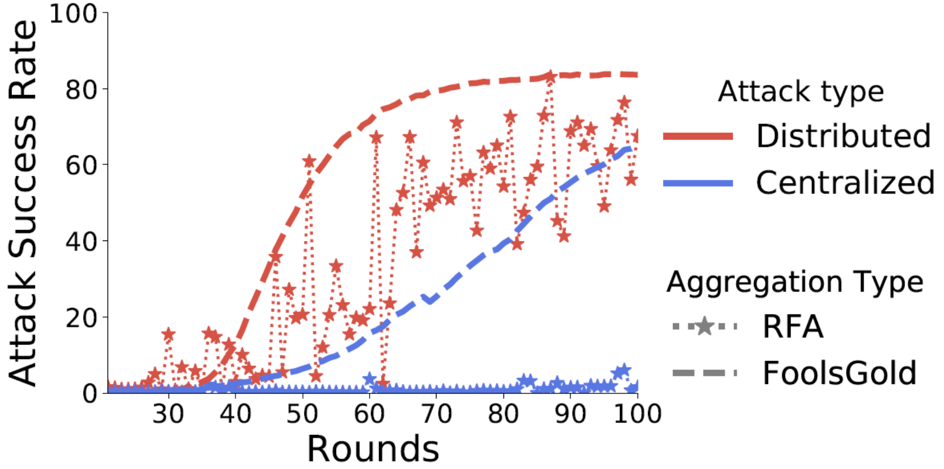
a) Training

b) Inference

Distributed Attack on Federated Learning



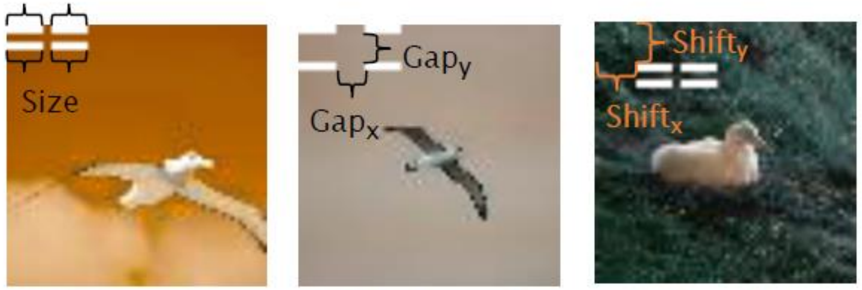
- Distributed backdoor attack is more effective, stealthier, and more resilient against "robust" aggregation



(d) Tiny-imagenet

Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. DBA: Distributed Backdoor Attacks against Federated Learning. ICLR 2020

More on Distributed Backdoor Attacks



(a) Trigger Size (b) Trigger Gap (c) Trigger Location

Figure 2: Trigger factors (size, gap and location) in backdoored images.

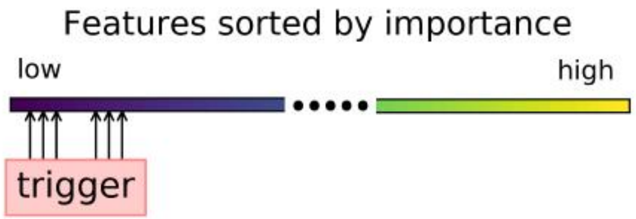


Figure 3: Trigger factor (feature importance ranking) in tabular data.

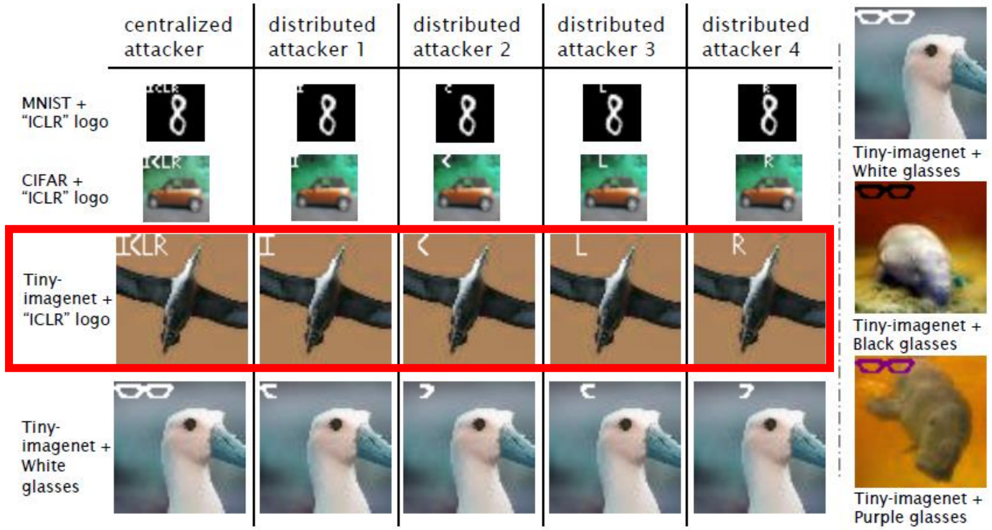
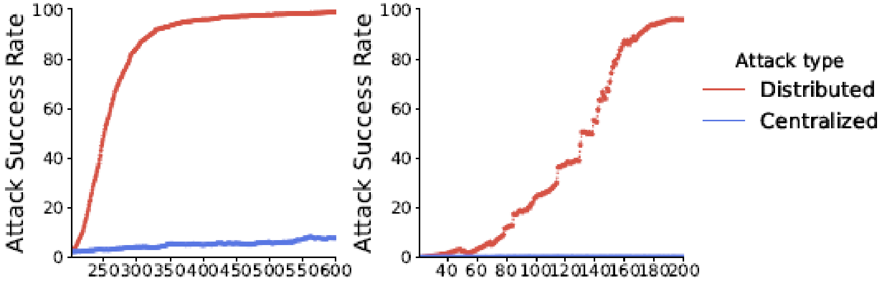


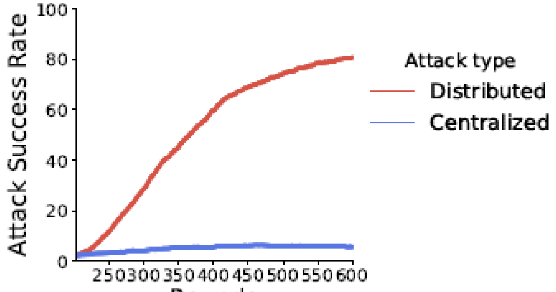
Figure 14: Examples of irregular shape triggers in image datasets

- Byzantine setting



(a) CIFAR (b) Tiny-imagenet

Figure 20: Multi-Krum



CIFAR

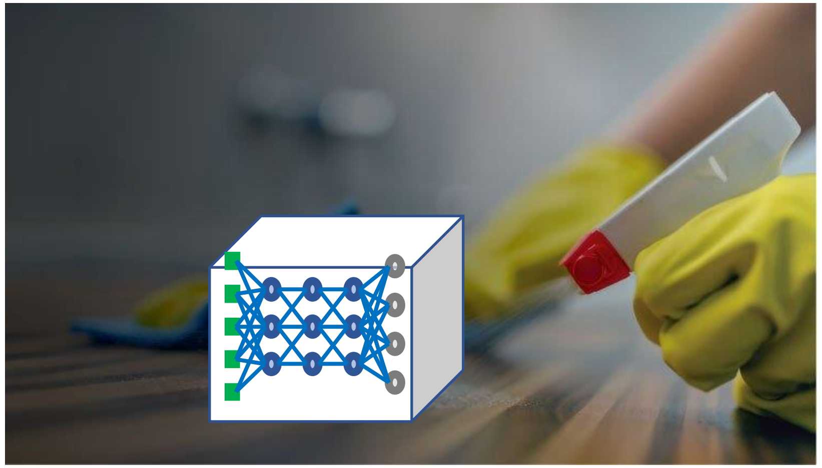
Figure 21: Bulyan

Why do we care? Model Sanitization!

- *I have an amazing ImageNet model which achieves 95% top-1 accuracy, and I make it publicly available by releasing the network architecture and trained model weights. Care to use it for your task?*

➢ Tempting … but *MLSS talk* makes me well educated. How do I know your model does not have any backdoor?

✓ Sanitize the model before using it (aka wear mask before you go out)



Yes! Using models from untrusted sources has risks of infection too!

Applications and Extensions based on Adversarial Attacks

Zeroth Order Optimization meets Black-box Attack

Black-box attack generation: an application of ZO optimization

- **A master problem:** $\min_{\boldsymbol{x}\in R^d} F(\boldsymbol{x}) = \sum_{i=1}^{n} f_i(\boldsymbol{x})$

f_i: **black-box/white-box** loss function at sample *i*

White-box attack generation

First-order optimization
e.g., stochastic gradient descent (SGD)

Black-box attack generation

Zeroth-order (ZO) optimization

unbiased: $E_i[\nabla f_i(\boldsymbol{x})] = \nabla F(\boldsymbol{x})$

Non-trivial

$\neq$

random gradient estimate : $\hat{\nabla} f_i(\boldsymbol{x}) = \frac{f_i(\boldsymbol{x}+\beta\boldsymbol{u}) - f_i(\boldsymbol{x})}{\beta}\boldsymbol{u}$

biased: $E_{i,u}\left[\hat{\nabla} f_i(\boldsymbol{x})\right] \neq \nabla F(\boldsymbol{x})$

SGD uses stochastic gradient $\nabla f_i(\boldsymbol{x})$

ZO uses gradient estimate $\hat{\nabla} f_i(\boldsymbol{x})$ via function queries

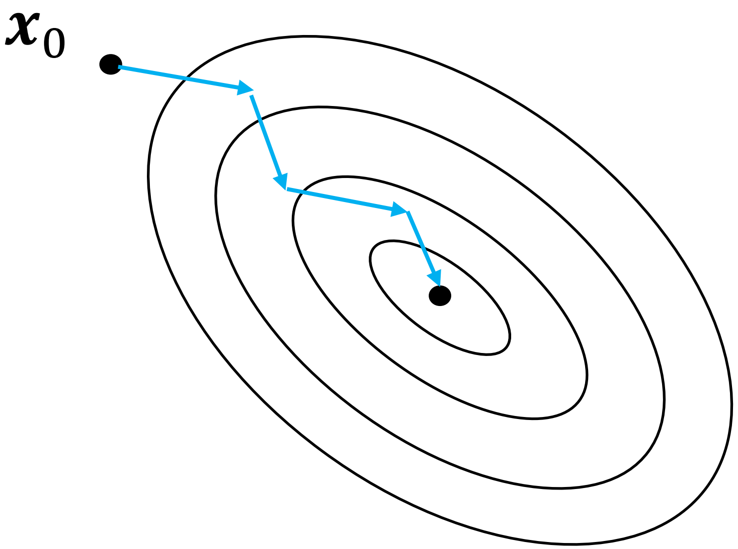
$$\boldsymbol{x}_k = \boldsymbol{x}_{k-1} - \alpha \nabla f_i(\boldsymbol{x}_{k-1}),\ k = 1,2,\ldots,T$$

$$\boldsymbol{x}_k = \boldsymbol{x}_{k-1} - \alpha \hat{\nabla} f_i(\boldsymbol{x}_{\boldsymbol{k}-\boldsymbol{1}}),\ k = 1,2,\ldots,T$$

$\alpha > 0$: step size

Zeroth-Order (ZO) Optimization

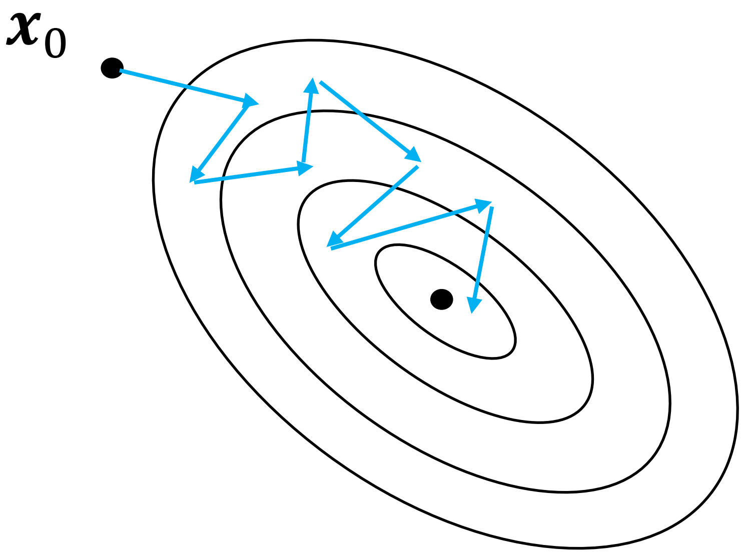
SGD (first order)



Convergence rate $E[\|\nabla F(\boldsymbol{x}_T)\|_2^2] = O(1/\sqrt{T})$

T is # of iterations

ZO-SGD



Convergence rate $E[\|\nabla F(\boldsymbol{x}_T)\|_2^2] = O(\sqrt{d}/\sqrt{T})$
[Duchi, et al., T-IT'15]

d is # of variables

Question: Better gradient estimate & ZO method with better convergence rate?

(Incomplete) Summary of Black-box Attack Methods

soft label = score based.   hard label = decision based.

- Transfer attack from white-box surrogate model [Papernot et. al.] (soft label)
- Zeroth-order optimization (ZO) based attack (feat. Convergence Guarantees)
  - ZO attack with gradient estimation [Chen et. al. AI Sec 2017] (soft label)
  - ZO-SVRG [Liu et. al. NeuRIPS 2018] (soft label)
  - ZO-Natural Evolution Strategy [Ilyas et. al. ICML 2018] (soft/hard label)
  - Input dimension reduction + ZO attack [Chen et. al. AAAI 2019] (soft label)
  - ZO-signSGD [Liu et. al. ICLR 2019] (soft label)
  - ZO-Natural Gradient Descent [Zhao et. al. AAAI 2019] (soft/hard label)
  - ZO-ADMM [Zhao et. al. ICCL 2019] (soft/hard label)
  - ZO-ADAM [Chen et. al. NeuRIPS 2019] (soft label)
  - ZO hard-label attack [Cheng et. al. ICLR 2019] (hard label)
  - Sign-OPT [Cheng et. al. ICLR 2020] (hard label)
- Bandit attack [Ilyas et. al. ICLR 2019] (soft label)
- Decision-based attack [Brendel et. al. ICLR 2018] (hard label)
- A lot more …

A Primer on Zeroth-Order Optimization in Signal Processing and Machine Learning

Sijia Liu, *Member, IEEE,* Pin-Yu Chen, *Member, IEEE,* Bhavya Kailkhura, *Member, IEEE,* Gaoyuan Zhang,
Alfred Hero, *Fellow, IEEE,* and Pramod K. Varshney, *Life Fellow, IEEE*

Survey paper: Liu, Chen, et al., "A Primer on Zeroth-Order Optimization in Signal Processing and Machine Learning", *IEEE Signal Processing Magazine*
https://arxiv.org/pdf/2006.06224.pdf

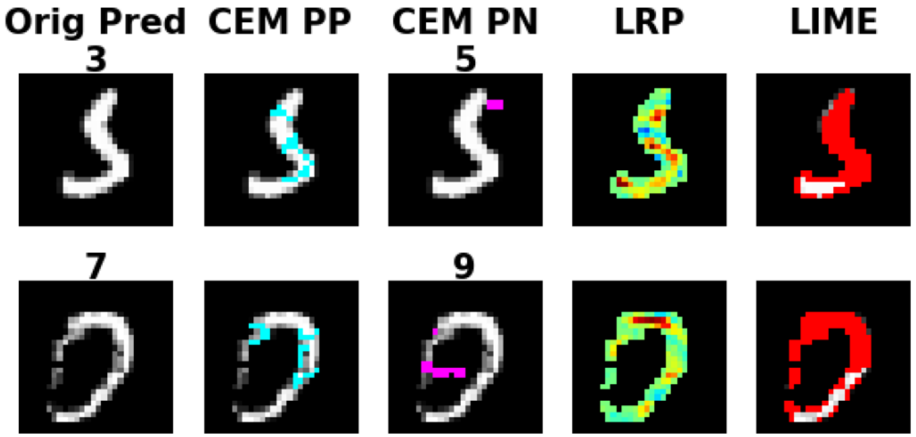
Applications and Extensions based on Adversarial Attacks

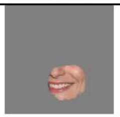
Adversarial Examples meets (Machine) Interpretation

Model Watermarking and Data Privacy

Generating Contrastive Explanations

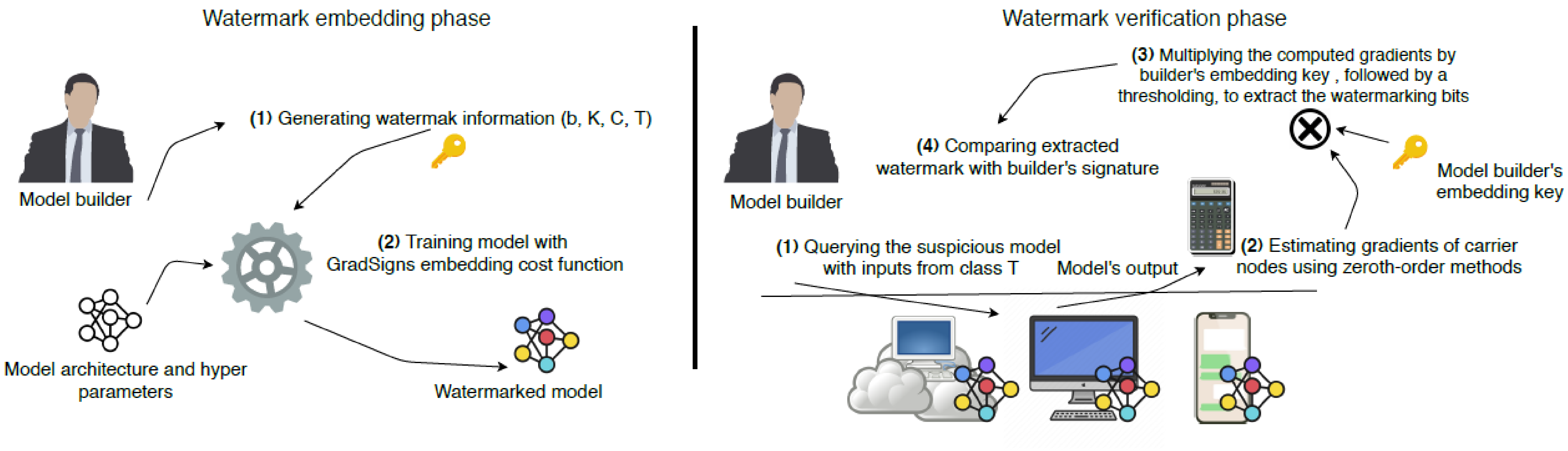
- ***Steve is the tall guy with long hair who does not wear glasses***
- Pertinent Positive (PP): minimally sufficient to be present to support the original classification
- Pertinent Negative (PN): necessarily absent to prevent changing the classification of the original image



Original Class Pred	yng, ml, smlg	yng, fml, smlg
Original		
Pert. Neg. Class Pred	old, ml, smlg	old, fml, smlg
Pertinent Negative		
Pert. Neg. Explanations	+gray hair	+oval face
Pertinent Positive		
LIME		
Grad-CAM		

Amit Dhurandhar*, Pin-Yu Chen*, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das, "Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives" NeurIPS 2018
Ronny Luss*, Pin-Yu Chen*, Amit Dhurandhar*, Prasanna Sattigeri*, Karthikeyan Shanmugam, and Chun-Chen Tu, "Generating Contrastive Explanations with Monotonic Attribute Functions" arxiv

Model Watermark Embedding and Extraction

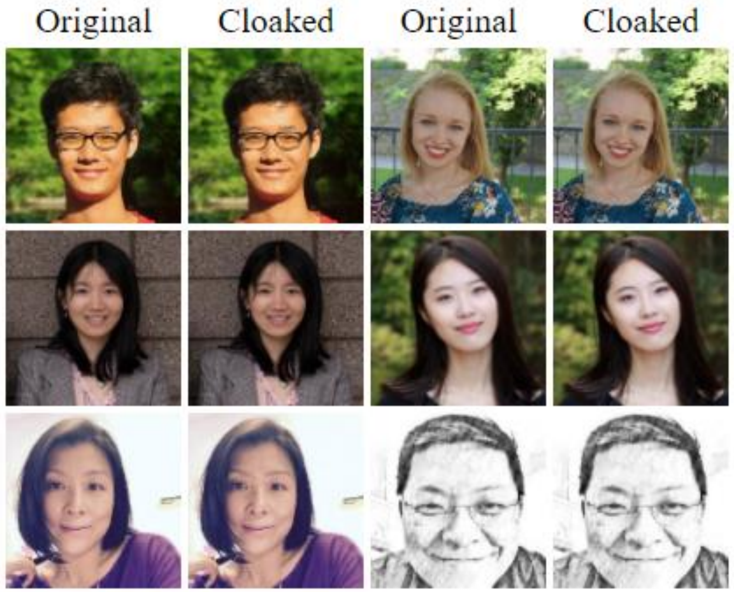


- Embed N-bit vector to a subset of dimension in input gradients
- Remote and black-box watermark extraction using gradient estimation

Omid Aramoon, Pin-Yu Chen, and Gang Qu. Don't Forget to Sign the Gradients! MLSyS 2021

Data Cloaking for Privacy

Image "Cloaking" for Personal Privacy

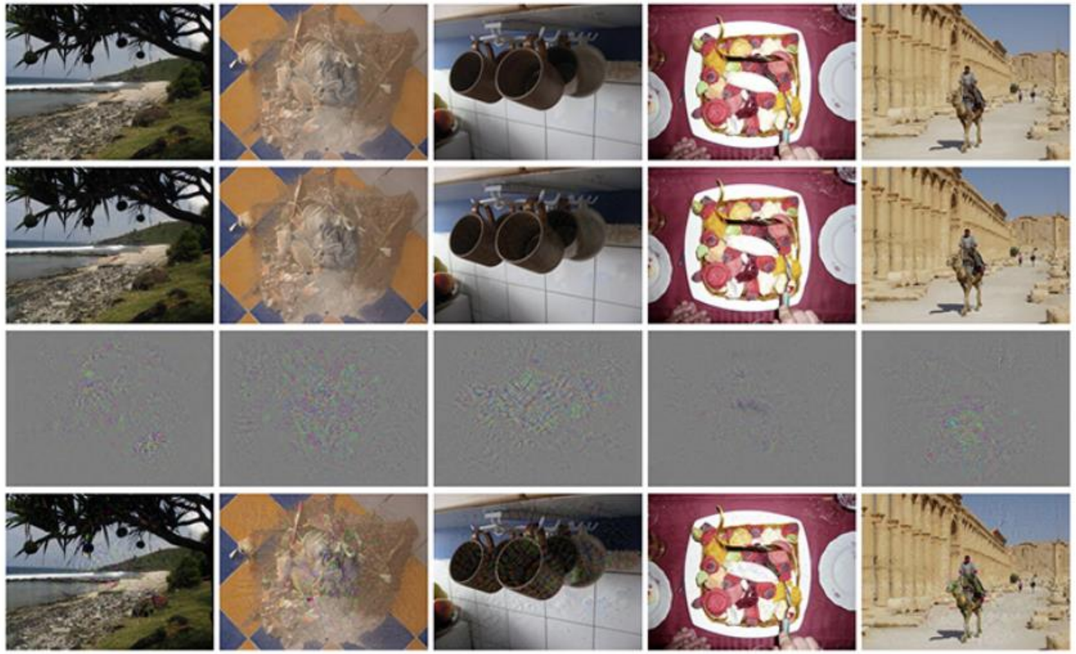


Shawn Shan†, PhD Student
Emily Wenger†, PhD Student
Jiayun Zhang, Visiting Student
Huiying Li, PhD Student
Haitao Zheng, Professor
Ben Y. Zhao, Professor
† *Project co-leaders and co-first authors*

- Email the Fawkes team
- Email us to join Fawkes mailing list for news on updates/changes.

SAND Lab
security, algorithms, networks and data

Using 'radioactive data' to detect if a dataset was used for training



The top row shows original images from the Holidays dataset and the second row shows the images with a radioactive mark (with PSNR=42dB). The third row shows the radioactive mark only, amplified by 5x. In the bottom row, this exaggerated mark is added to the original images for visualization purposes, which amounts to a 14dB amplification of the additive noise.

https://sandlab.cs.uchicago.edu/fawkes/

https://ai.facebook.com/blog/using-radioactive-data-to-detect-if-a-data-set-was-used-for-training/

More Interesting Applications

Ad-versarial: Perceptual Ad-Blocking meets Adversarial Machine Learning

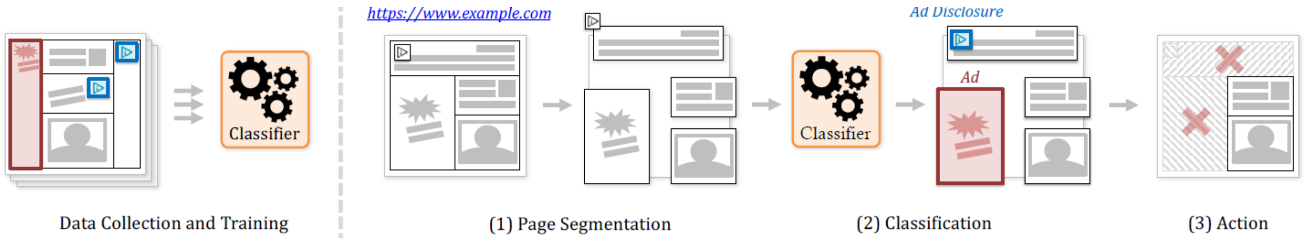
Florian Tramèr
Stanford University

Pascal Dupré
CISPA

Gili Rusak
Stanford University

Giancarlo Pellegrino
Stanford University & CISPA

Dan Boneh
Stanford University



(a) **Original Page:** two ads are detected.

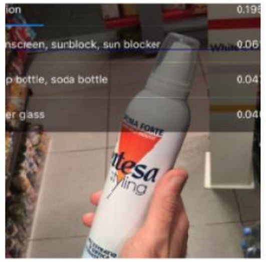
(b) **Attack C4-U:** The publisher overlays a transparent mask over the full page to evade the ad-blocker.

(c) **Attack C4-U':** The publisher overlays a mask on the page to generate unreasonably large boxes and disable the ad-blocker.

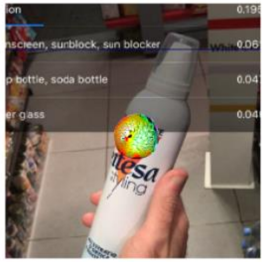
(d) **Attack C1-U:** The publisher adds an opaque footer to detect an ad-blockers that blocks the honeypot element (bottom-left).

Shoplifting Smart Stores Using Adversarial Machine Learning

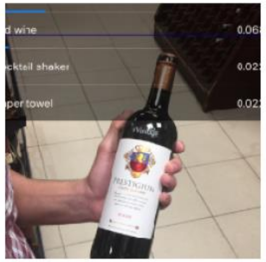
Mohamed Nassar, Abdallah Itani, Mahmoud Karout,
Mohamad El Baba, Omar Al Samman Kaakaji
Department of Computer Science
Faculty of Arts and Sciences
American University of Beirut (AUB)
Beirut, Lebanon



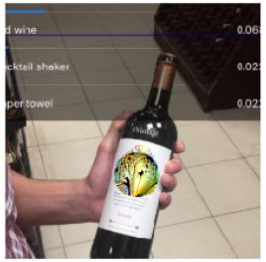
(c) Hair spray



(d) Hair spray as an orange (confidence = 66%)



(e) Wine bottle

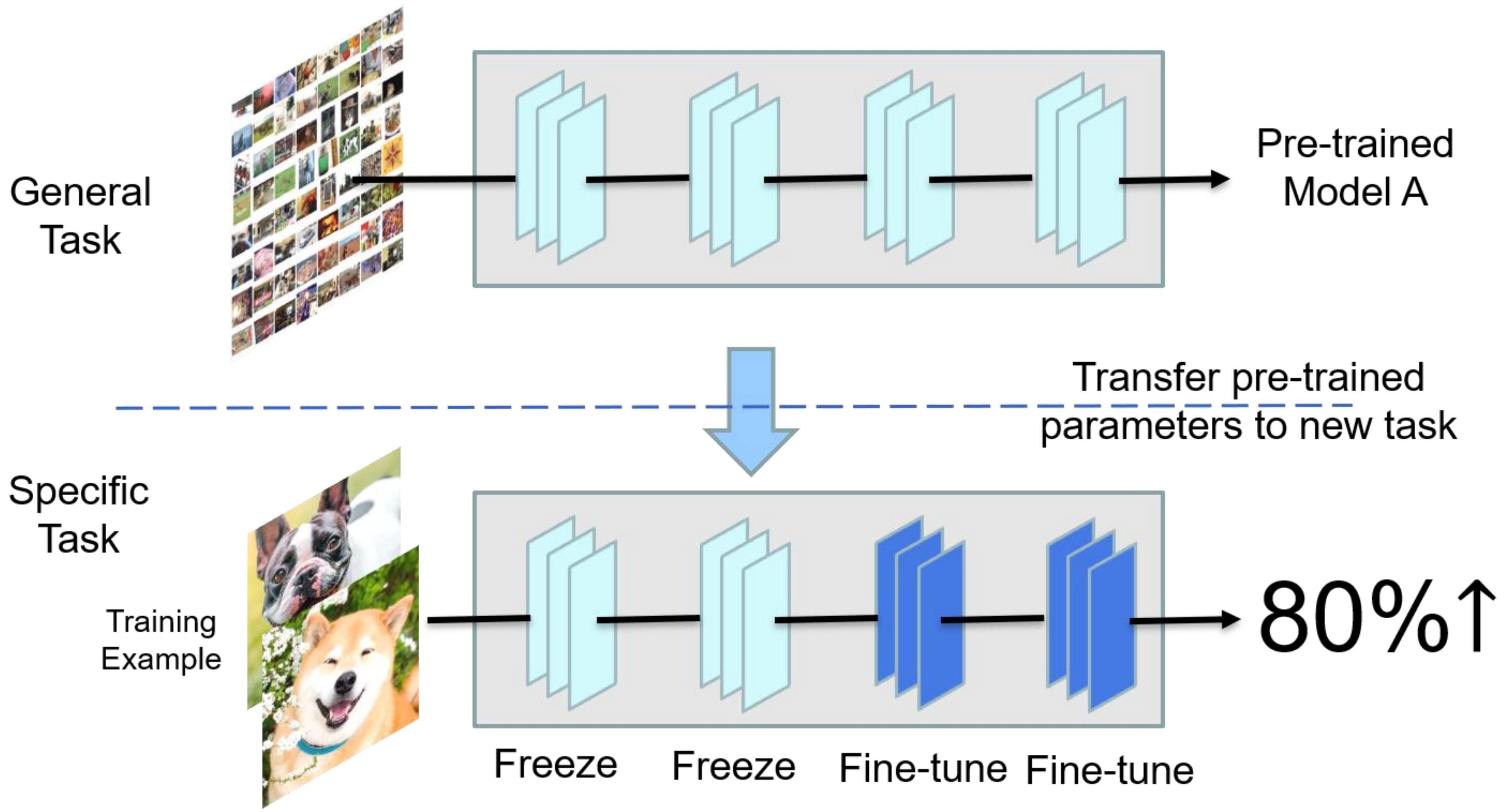


(f) Wine bottle as a banana (confidence = 78%)

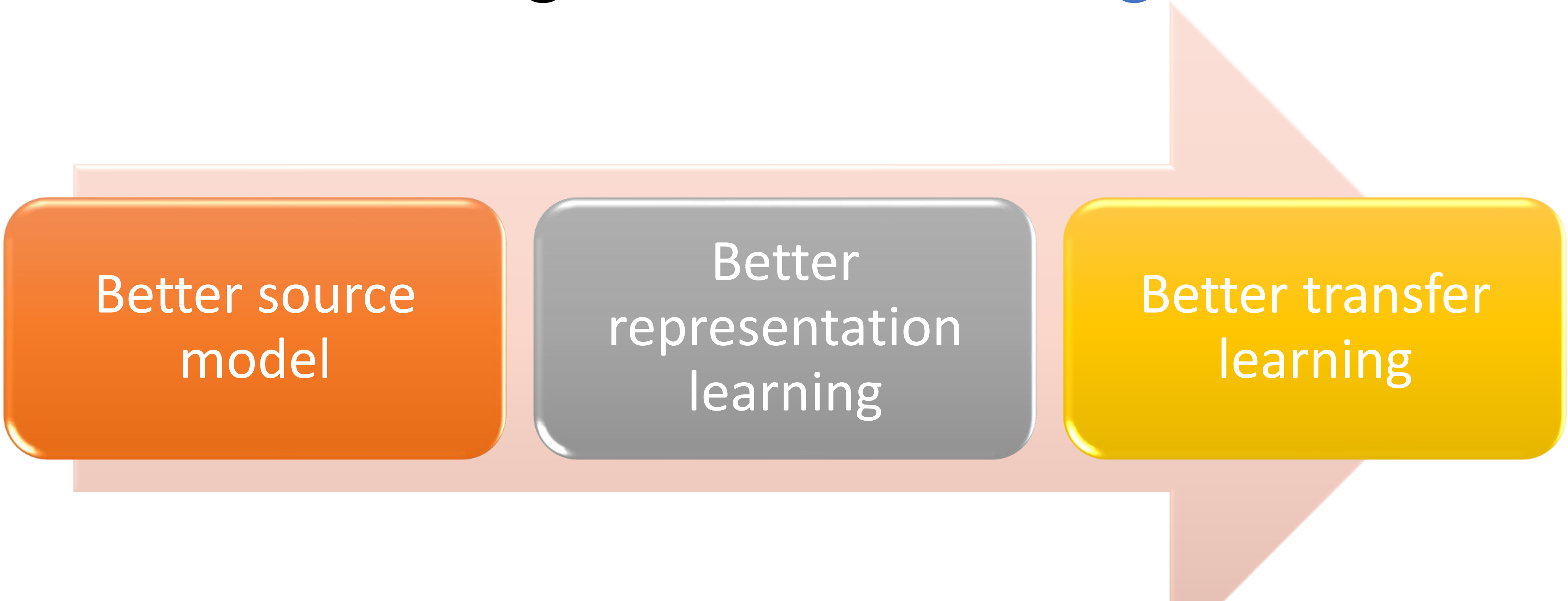
Q&A for Part I

Model Reprogramming: Adversarial ML for Good

Transfer Learning via Fine-Tuning



Transfer Learning without Knowing?

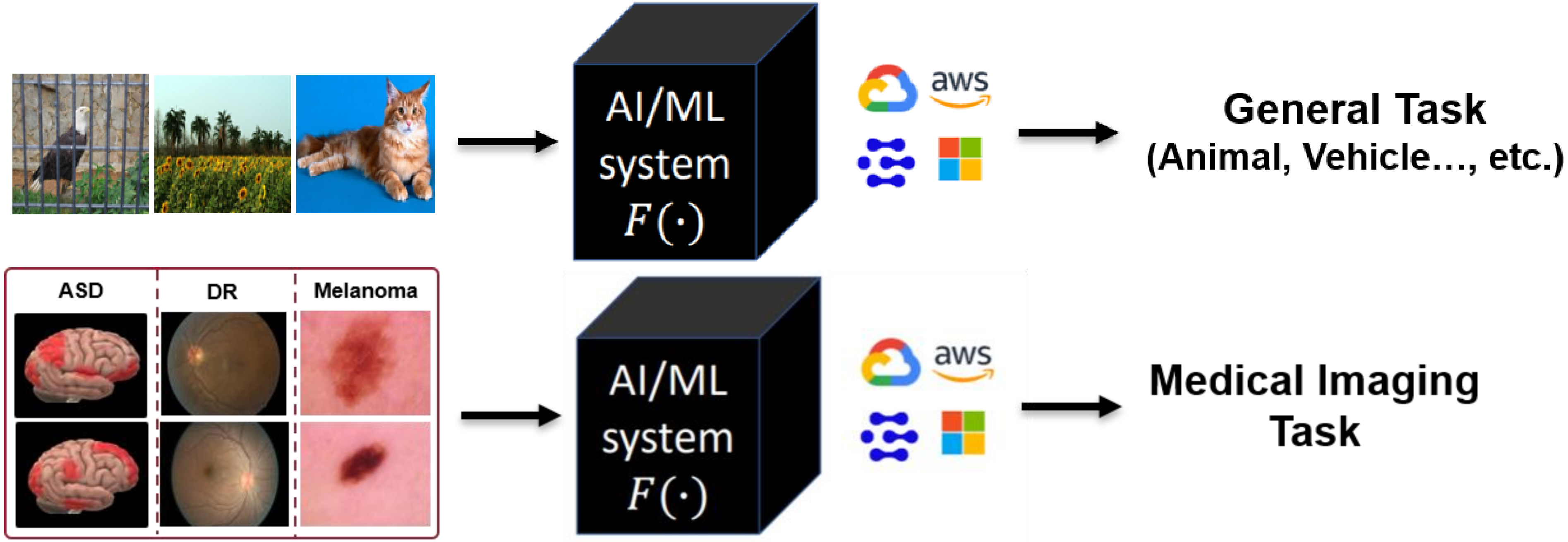


- Are we able to do transfer learning on the “best” model?

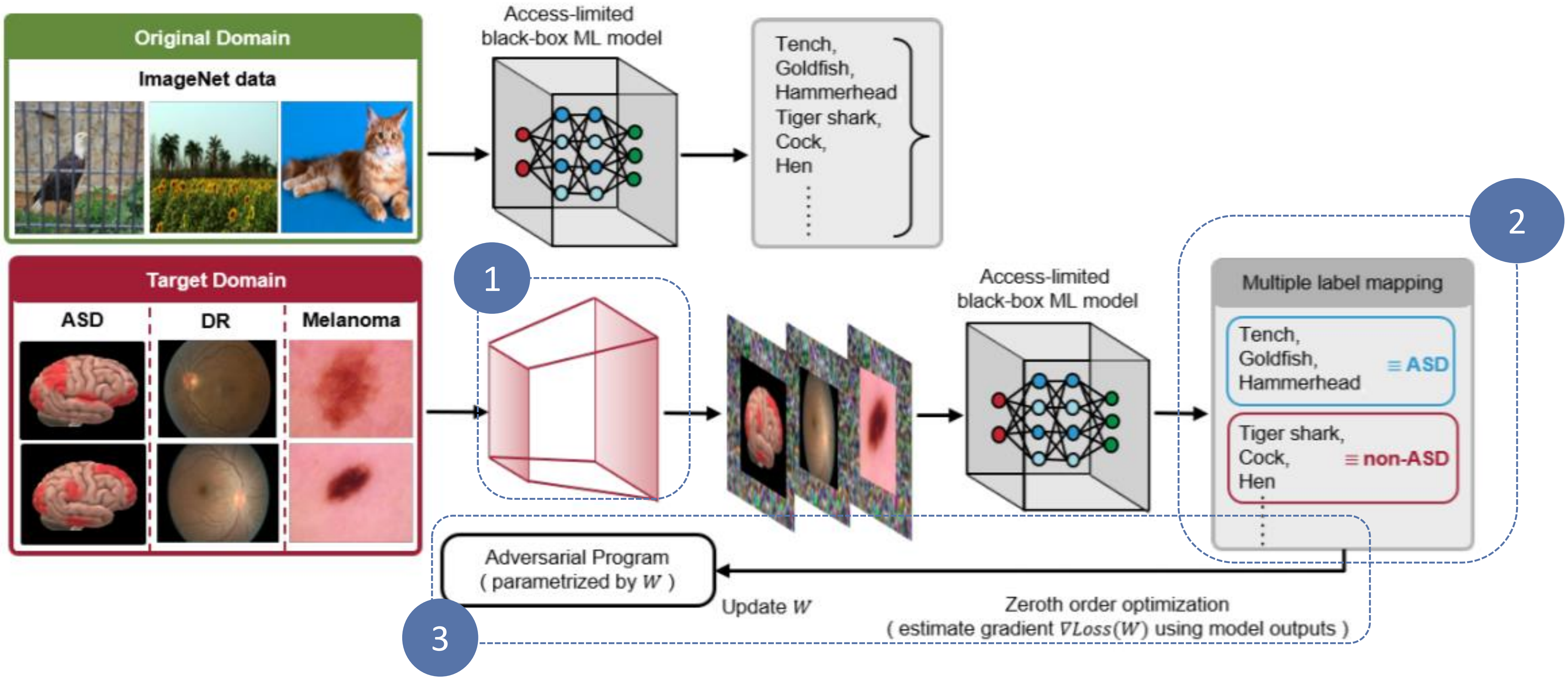
➢ Not really, especially when they are black-box models

Black-box Adversarial Reprogramming (BAR)

- Reprogram powerful but black-box models for transfer learning (w/o fine-tuning) – *teach old dog new tricks*
- Appealing for cross-domain and data-limited transfer learning



Black-box Adversarial Reprogramming (BAR):
Data-Efficient Transfer Learning



Yun-Yun Tsai, Pin-Yu Chen, Tsung-Yi Ho. Transfer Learning without Knowing: Reprogramming Black-box Machine Learning Models with Scarce Data and Limited Resources. ICML 2020

Credit: Yun-Yun Tsai@NTHU

Problem Formulation

- Given a black-box model:

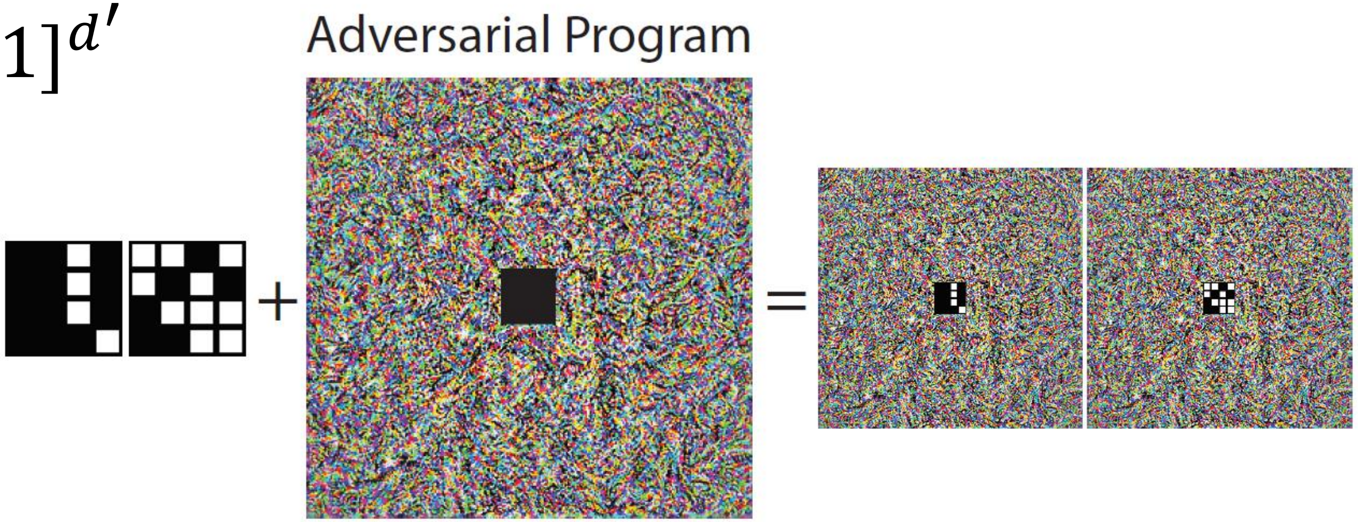
$$F\ : \mathcal{X} \rightarrow \mathbb{R}^K,$$

where $\mathcal{X} \in [-1, 1]^d$ and $F(x) = [F_1(x), F_2(x), \ldots, F_K(x)] \in \mathbb{R}^K$

- Given the set of data from the target domain by:

$\{T_i\}_{i=1}^{n}$, where $T_i \in [-1, 1]^{d'}$ and $d' < \ d$

- Output: Optimal adversarial program with parameters W.



Elsayed, Gamaleldin F., Ian Goodfellow, and Jascha Sohl-Dickstein. "Adversarial reprogramming of neural networks." ICLR 2019

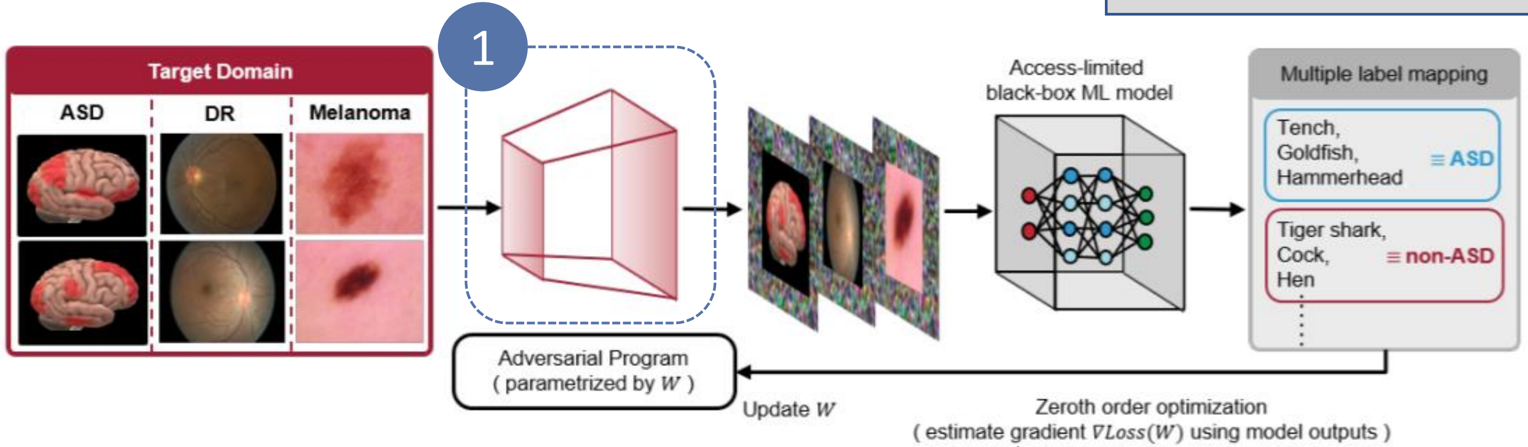
Adversarial Program Function

- The transformed data sample for BAR is defined as:

$$\tilde{X}_i = \{T_i\}_{padding} + P, \text{ and } P = tanh(W \odot M)$$

Universal trainable perturbation (aka Trigger!)

Trainable parameters: $W \in \mathbb{R}^d$



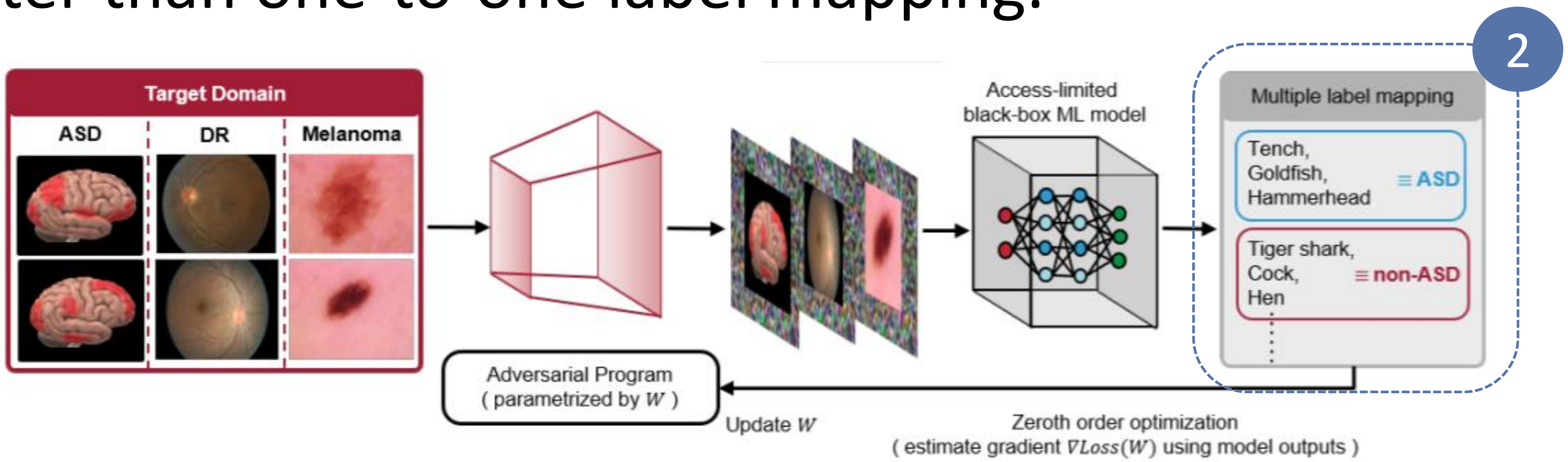
Gamaleldin F. Elsayed, Ian Goodfellow, Jascha Sohl-Dickstein. Adversarial Reprogramming of Neural Networks. ICLR 2019

Multi-label Mapping (Random)

- We use the notation $h_j\ (\cdot)$ to denote $m\ to\ 1$ mapping function. For example,

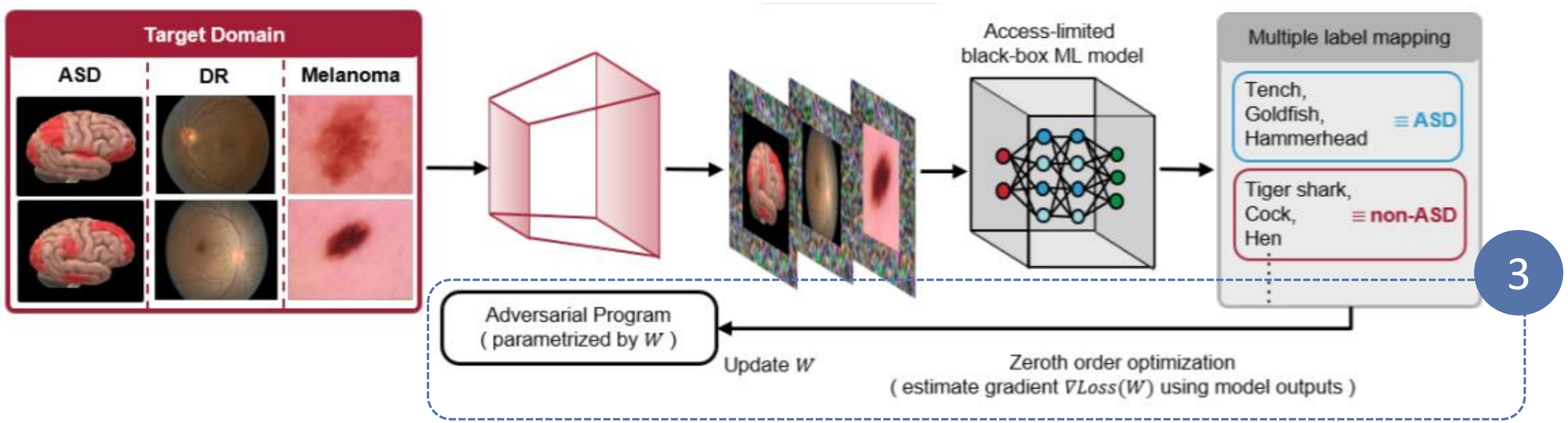
$$h_{ASD}\big(F(X)\big) = \frac{F_{Tench}(X) + F_{Goldenfish}(X)\ +\ F_{Hammerhead}(X)}{3}$$

- We find that multiple-source-labels to one target-label mapping better than one-to-one label mapping.



Training Loss Function

- We aim to maximize the probability of $p_t = P\left(h_j\left(y_{target}\right)\middle|X_{target}\right)$
- We use focal loss empirically as it can further improve the performance of AR/BAR over cross entropy. $L_{focal}(p_t) = -\omega(1-p_t)^{\gamma}log(p_t)$
- ZO optimization for learning W in BAR : $W_{t+1} = W_t - \alpha_t \cdot \hat{\nabla} L(W_t)$



Lin et al. Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision, pp. 2980–2988, 2017.

Experimental Results

- **Autism Spectrum Disorder Classification (2 classes)**
  - We use Autism Brain Imaging Data Exchange (ABIDE) database.
  - It contains 503 individuals suffering from ASD and 531 non-ASD samples.
  - The data sample is a 200×200 brain-regional correlation graph of fMRI measurements.

Model	Accuracy	Sensitivity	Specificity
Resnet 50 (AR)	72.99%	73.03%	72.13%
Resnet 50 (BAR)	70.33%	69.94%	72.71%
Train from scratch	50.96%	50.13%	52.34%
Transfer Learning (finetuned)	52.88%	54.13%	53.50%
Incept.V3 (AR)	72.30%	71.94%	74.71%
Incept.V3 (BAR)	70.10%	69.40%	70.00%
Train from scratch	49.80%	50.40%	51.55%
Transfer Learning (finetuned)	50.10%	51.23%	47.42%
SOTA 1. (Heinsfeld et al., 2018)	65.40%	69.30%	61.10%
SOTA 2. (Eslami et al., 2019)	69.40%	66.40%	71.30%

Eslami et al. Asd-diagnet: A hybrid learning approach for detection of autism spectrum disorder using fmri data. Frontiers in Neuroinformatics, 13, Nov 2019.

Experimental Results

- Melanoma Detection (7 classes)
  - The target-domain dataset is from the International Skin Imaging Collaboration (ISIC) dataset.
  - The performance of SOTA is 78.65%, which uses specifically designed data augmentation with finetuning on Densenet.

Model	From Stratch	Finetuning	AR	BAR
Resnet 50	59.01%	76.90%	82.05%	81.71%
Incept.V3	52.91%	58.63%	82.01%	80.20%
Densenet 121	52.28%	58.88%	80.76%	78.33%

Li, et al. Skin lesion analysis towards melanoma detection via end-to-end deep learning of convolutional neural networks. arXiv preprint arXiv:1807.08332, 2018.

Experimental Results

- Reprogramming Microsoft Custom Vision API:
  - This API allows user uploading labeled datasets and training an ML model for prediction.
  - The model is unknown to end user.
  - We use this API and train a traffic sign image recognition model (43 classes) using GTSRB dataset.

Orig. Task to New Task	q	# of query	Accuracy	Cost
Traffic sign classification	1	1.86k	48.15%	\$3.72
to	5	5.58k	62.34%	\$11.16
ASD	10	10.23k	**67.80%**	\$20.46

V2S: Reprogramming Human Acoustic Models for (Univariate) Time-Series Classification

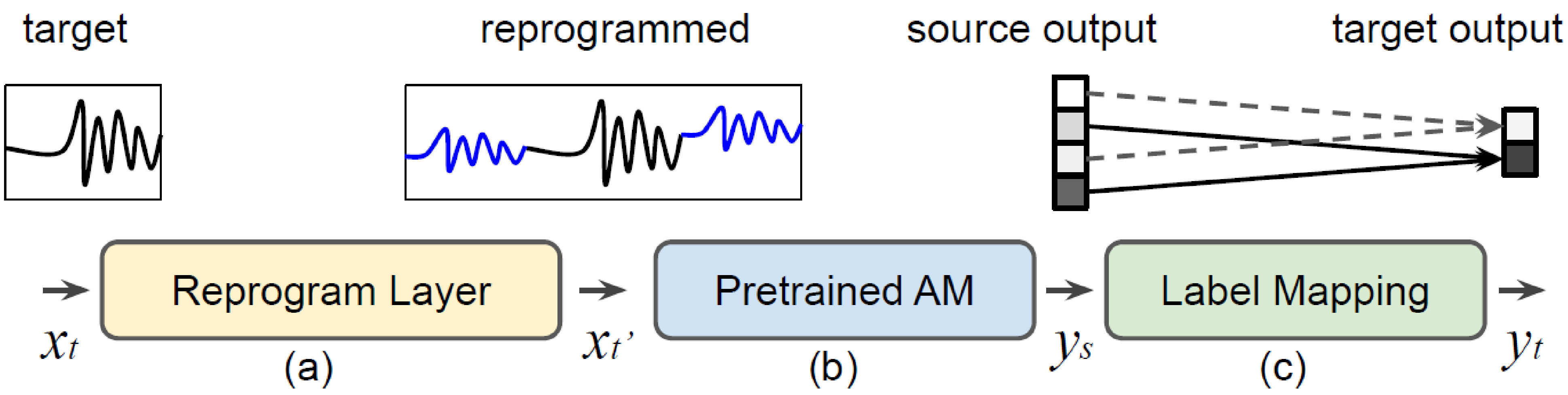


Figure 1: Schematic illustration of the proposed Voice2Series (V2S) framework: (a) trainable reprogram layer; (b) pre-trained acoustic model (AM); (c) source-target label mapping function.

Chao-Han Huck Yang, Yun-Yun Tsai and Pin-Yu Chen "**Voice2Series: Reprogramming Acoustic Models for Time Series Classification**," *ICML 2021*

V2S Algorithm and Implementation

Algorithm 1 Voice to Series (V2S) Reprogramming

1: **Inputs**: Pre-trained acoustic model f_S, V2S loss L in (3), target domain training data $\{x_t^{(i)}, y_t^{(i)}\}_{i=1}^{n}$, mask function M, multi-label mapping function $h(\cdot)$, maximum number of iterations T, initial learning rate α

2: **Output**: Optimal reprogramming parameters θ^*

3: Initialize θ randomly; set $t = 0$

4: **#Generate reprogrammed data input**
$\mathcal{H}(x_t^{(i)}; \theta) = \text{Pad}(x_t^{(i)}) + M \odot \theta, \forall\, i = \{1, 2, \ldots, n\}$

5: **#Compute V2S loss L from equation (3)**
$L(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \log P(y_t^{(i)}) | f_S(\mathcal{H}(x_t^{(i)}); \theta))$

6: **#Solve reprogramming parameters**
Use ADAM optimizer to solve for θ^* based on $L(\theta)$

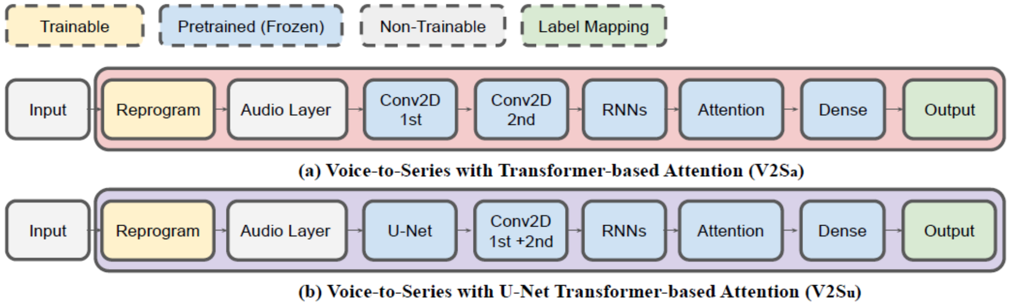


Figure 2: V2S architectures: (a) V2S$_a$ (de Andrade et al., 2018) and (b) V2S$_u$ (Yang et al., 2020).

V2S Outperforms SOTA on 20/30 UCR Datasets!

Table 2. Performance comparison of test accuracy (%) on 30 UCR time series classification datasets (Dau et al., 2019). Our proposed $V2S_a$ outperforms or ties with the current SOTA results (discussed in Section 5.3) on 20 out of 30 datasets.

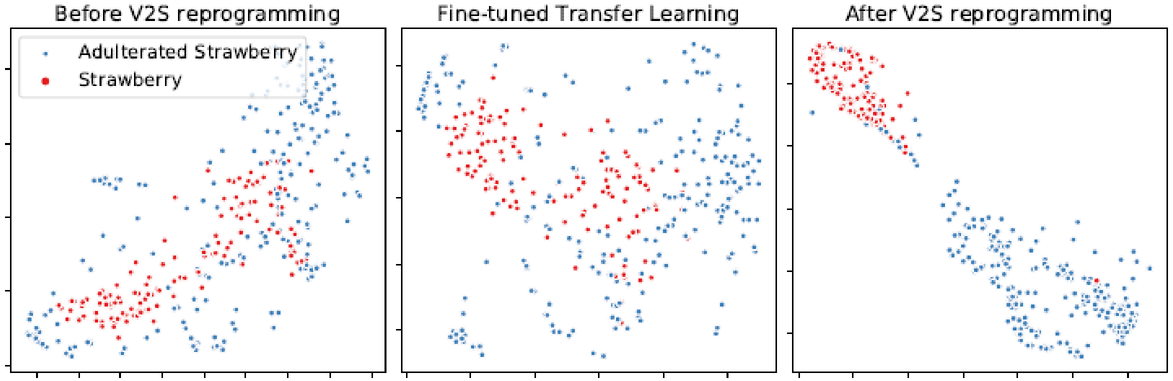
Dataset	Type	Input size	Train. Data	Class	SOTA	$V2S_a$	$V2S_u$	TF_a
Coffee	SPECTRO	286	28	2	**100**	**100**	**100**	53.57
DistalPhalanxTW	IMAGE	80	400	6	**79.28**	79.14	75.34	70.21
ECG 200	ECG	96	100	2	90.9	**100**	**100**	**100**
ECG 5000	ECG	140	500	5	**94.62**	93.96	93.11	58.37
Earthquakes	SENSOR	512	322	2	76.91	**78.42**	76.45	74.82
FordA	SENSOR	500	2500	2	96.44	**100**	**100**	**100**
FordB	SENSOR	500	3636	2	92.86	**100**	**100**	**100**
GunPoint	MOTION	150	50	2	**100**	96.67	93.33	49.33
HAM	SPECTROM	431	109	2	**83.6**	78.1	71.43	51.42
HandOutlines	IMAGE	2709	1000	2	**93.24**	**93.24**	91.08	64.05
Haptics	MOTION	1092	155	5	51.95	**52.27**	50.32	21.75
Herring	IMAGE	512	64	2	**68.75**	**68.75**	64.06	59.37
ItalyPowerDemand	SENSOR	24	67	2	97.06	**97.08**	96.31	97
Lightning2	SENSOR	637	60	2	86.89	**100**	**100**	**100**
MiddlePhalanxOutlineCorrect	IMAGE	80	600	2	72.23	**83.51**	81.79	57.04
MiddlePhalanxTW	IMAGE	80	399	6	58.69	**65.58**	63.64	27.27
Plane	SENSOR	144	105	7	**100**	**100**	**100**	9.52
ProximalPhalanxOutlineAgeGroup	IMAGE	80	400	3	88.09	**88.78**	87.8	48.78
ProximalPhalanxOutlineCorrect	IMAGE	80	600	2	**92.1**	91.07	90.03	68.38
ProximalPhalanxTW	IMAGE	80	400	6	81.86	**84.88**	83.41	35.12
SmallKitchenAppliances	DEVICE	720	375	3	**85.33**	83.47	74.93	33.33
SonyAIBORobotSurface	SENSOR	70	20	2	**96.02**	**96.02**	91.71	34.23
Strawberry	SPECTRO	235	613	2	**98.1**	97.57	91.89	64.32
SyntheticControl	SIMULATED	60	300	6	**100**	98	99	49.33
Trace	SENSOR	271	100	4	**100**	**100**	**100**	18.99
TwoLeadECG	ECG	82	23	2	**100**	96.66	97.81	49.95
Wafer	SENSOR	152	1000	2	99.98	**100**	**100**	100
WormsTwoClass	MOTION	900	181	2	83.12	**98.7**	90.91	57.14
Worms	MOTION	900	181	5	80.17	**83.12**	80.34	42.85
Wine	SPECTRO	234	57	2	**92.61**	90.74	90.74	50
Mean accuracy (↑)	-	-	-	-	88.02	**89.86**	87.92	56.97
Median accuracy (↑)	-	-	-	-	92.36	**94.99**	91.40	53.57
MPCE (mean per class error) (↓)	-	-	-	-	2.09	**2.01**	2.10	48.34

Why and When Model Reprogramming Works? (No, it's not about knowledge transfer)

Theorem 1: Let δ^* denote the learned additive input transformation for reprogramming (Assumption 4). The population risk for the target task via reprogramming a K-way source neural network classifier $f_{\mathcal{S}}(\cdot) = \eta(z_{\mathcal{S}}(\cdot))$, denoted by $\mathbb{E}_{\mathcal{D}_{\mathcal{T}}}[\ell_{\mathcal{T}}(x_t + \delta^*, y_t)]$, is upper bounded by

$$\mathbb{E}_{\mathcal{D}_{\mathcal{T}}}[\ell_{\mathcal{T}}(x_t + \delta^*, y_t)] \leq \underbrace{\epsilon_{\mathcal{S}}}_{\text{source risk}}$$

$$+ 2\sqrt{K} \cdot \underbrace{\mathcal{W}_1(\mu(z_{\mathcal{S}}(x_t + \delta^*)), \mu(z_{\mathcal{S}}(x_s)))_{x_t \sim \mathcal{D}_{\mathcal{T}},\, x_s \sim \mathcal{D}_{\mathcal{S}}}}_{\text{representation alignment loss via reprogramming}}$$



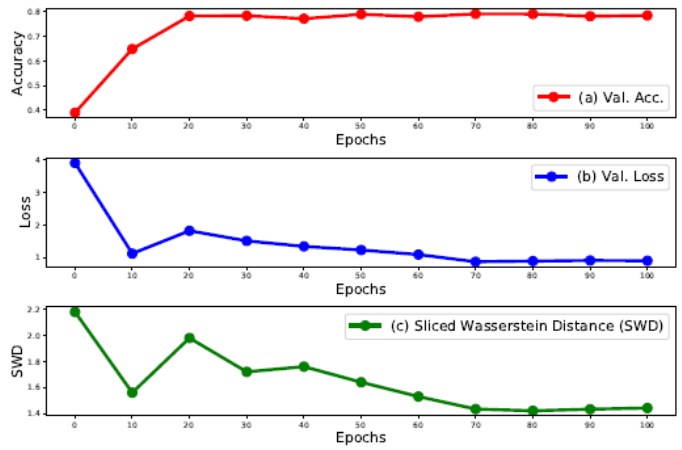


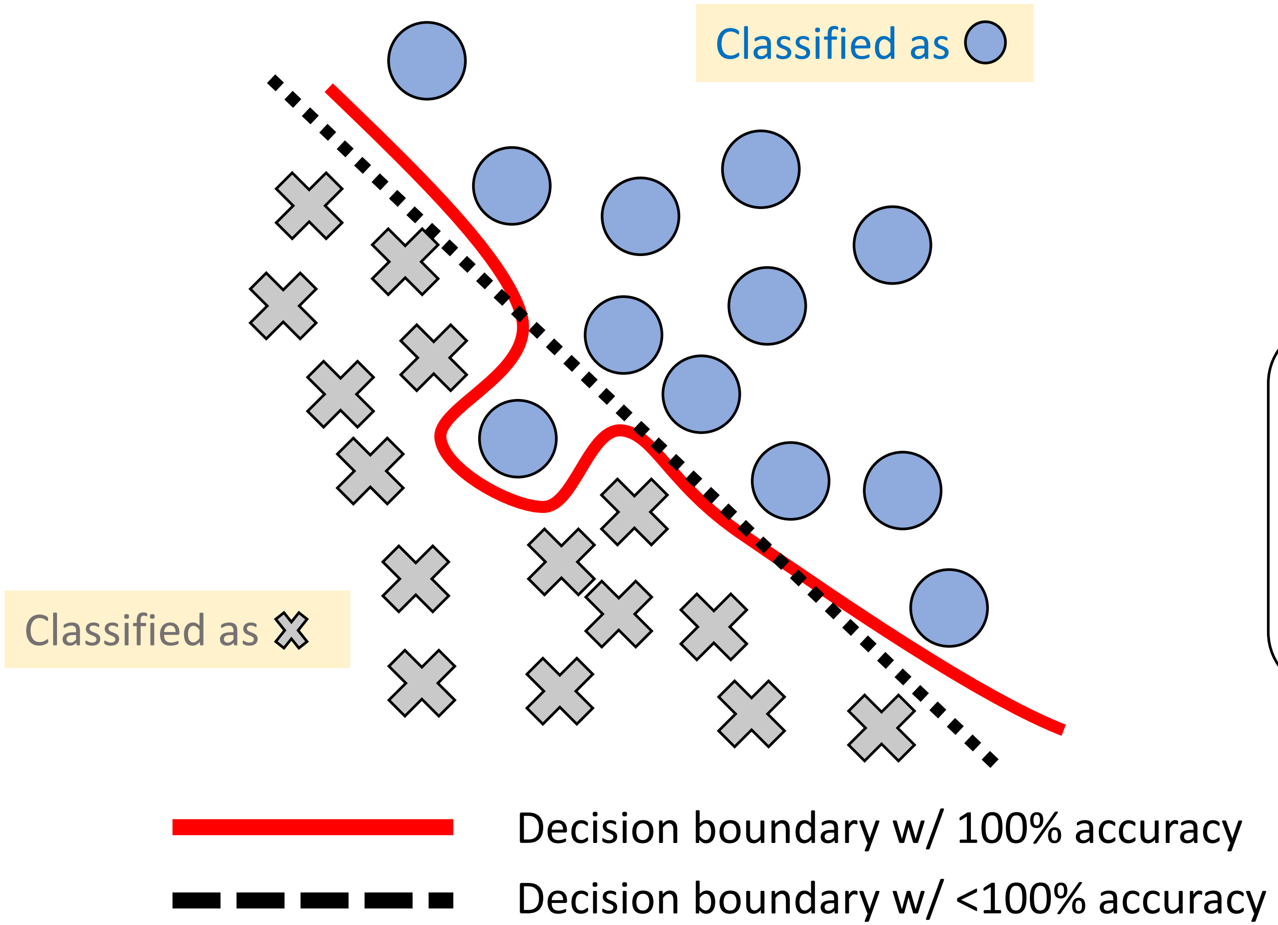
Figure 3: Training-time reprogramming analysis using V2S$_a$ and DistalPhalanxTW dataset (Davis, 2013). All values are averaged over the training set. The rows are (a) validation (test) accuracy, (b) validation loss, and (c) sliced Wasserstein distance (SWD) (Kolouri et al., 2018).

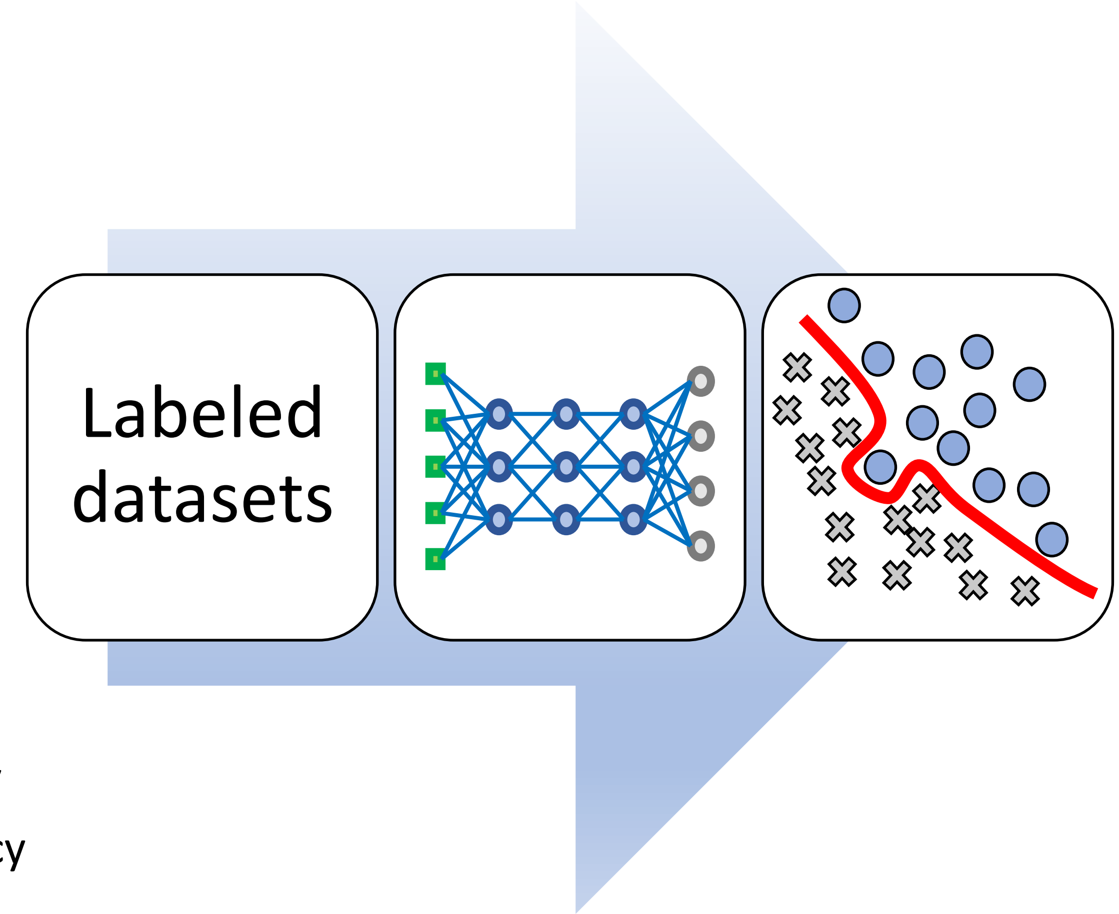
Table 3: Validation loss (Loss$_{\mathcal{S}}$) of the source task (GSCv2 voice dataset (Warden, 2018)) and mean/median Sliced Wasserstein Distance (SWD) of all training sets in Table 2.

Model	Loss$_{\mathcal{S}}$	Mean SWD	Median SWD
V2S$_a$	**0.1709**	**1.829**	**1.943**
V2S$_u$	0.1734	1.873	1.977

Adversarial Defenses: empirically v.s. provable robustness

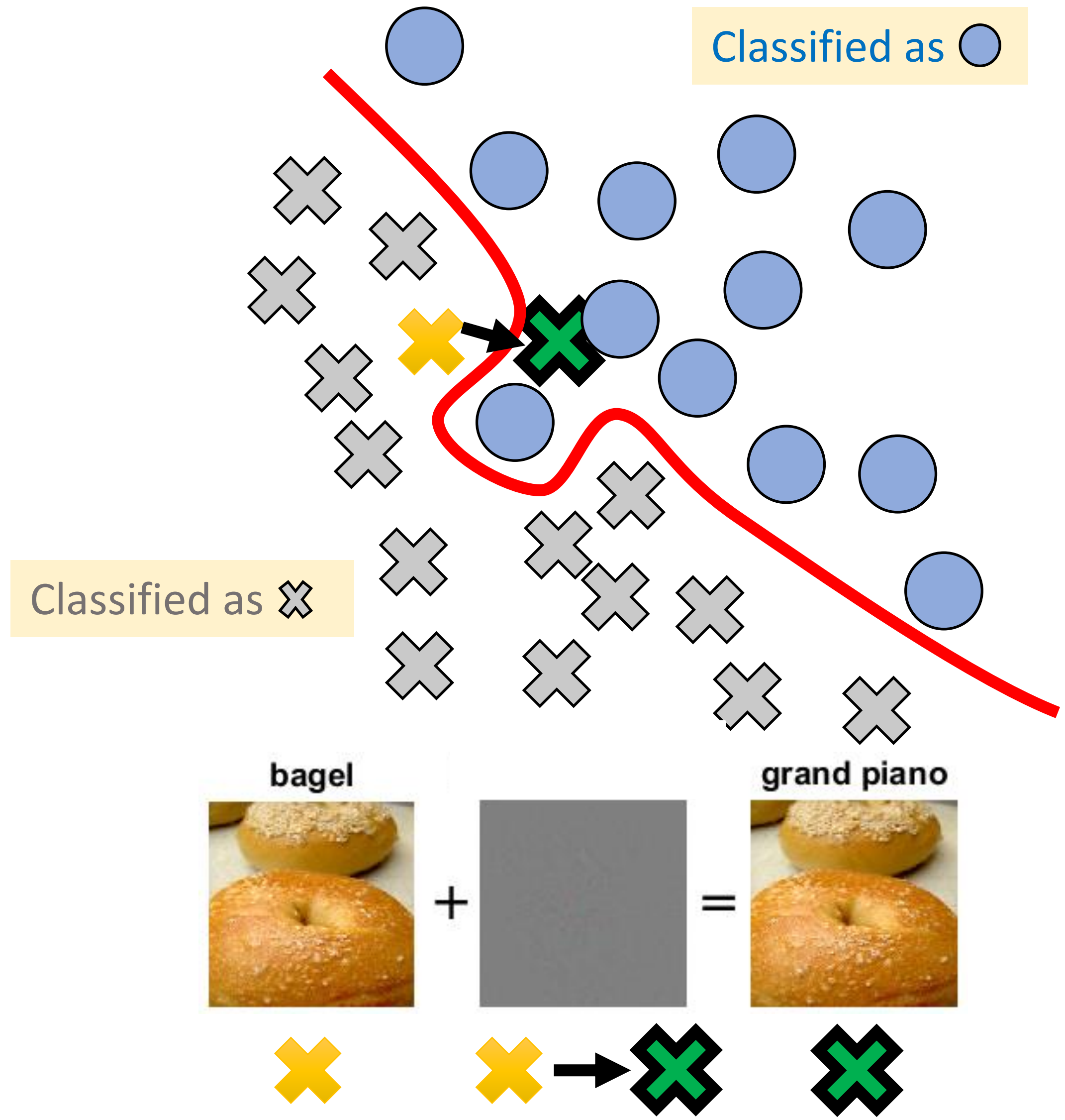
Learning to classify is all about drawing a line

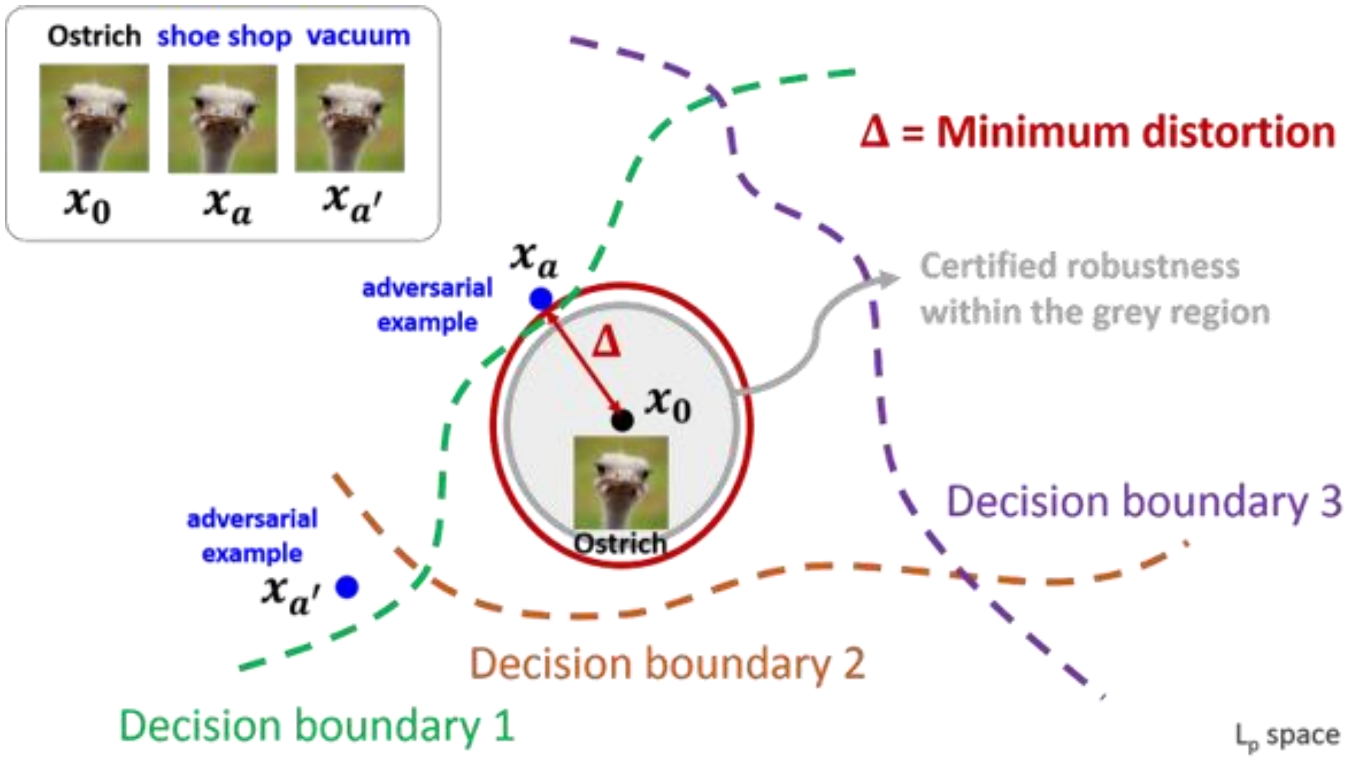




Source: Paishun Ting

Connecting adversarial examples to model robustness



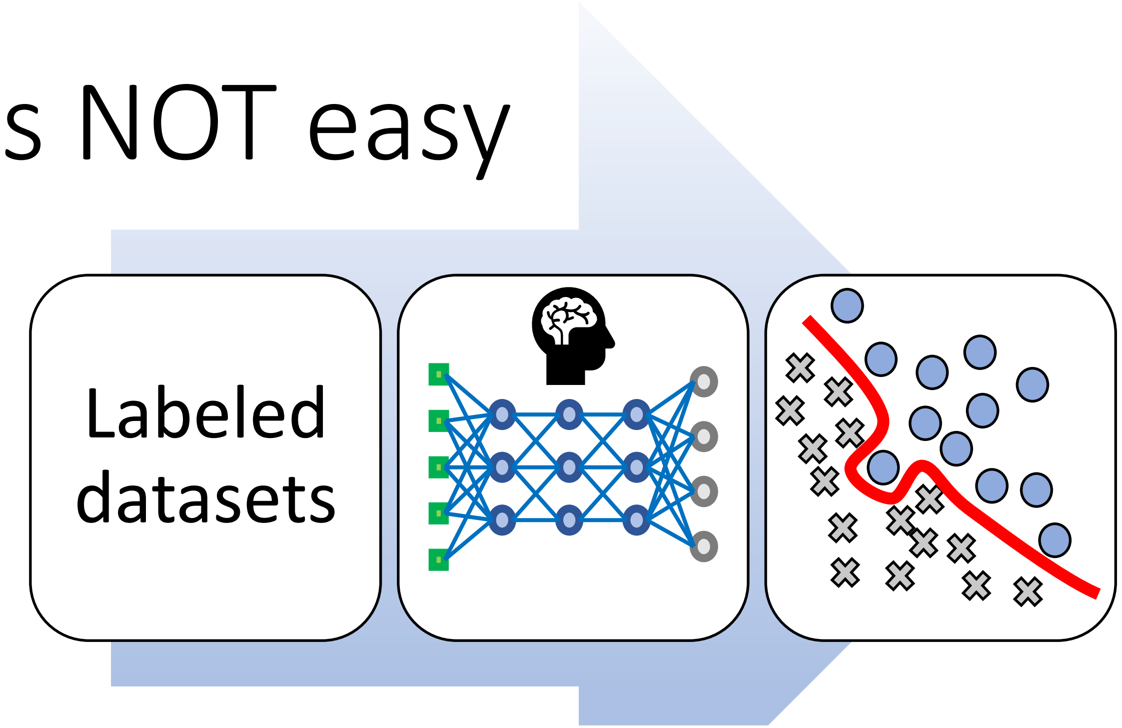


- Robustness evaluation: how close a refence input is to the (closest) decision boundary

Source:Paishun-Ting, Tsui-Wei Weng

Learning a robust model is NOT easy

- We still don’t fully understand how neural nets learn to predict
- ❑ calling for interpretable AI
- Training data could be noisy and biased
- ❑ calling for robust and fair AI
- Neural network architecture could be redundant and leading to vulnerable spots
- ❑ calling for efficient and secure AI model
- Need for human-like machine perception and understanding
- ❑ calling for bio-inspired AI model
- Attacks can also benefit and improve upon the progress in AI
- ❑ calling for attack-independent evaluation



Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods

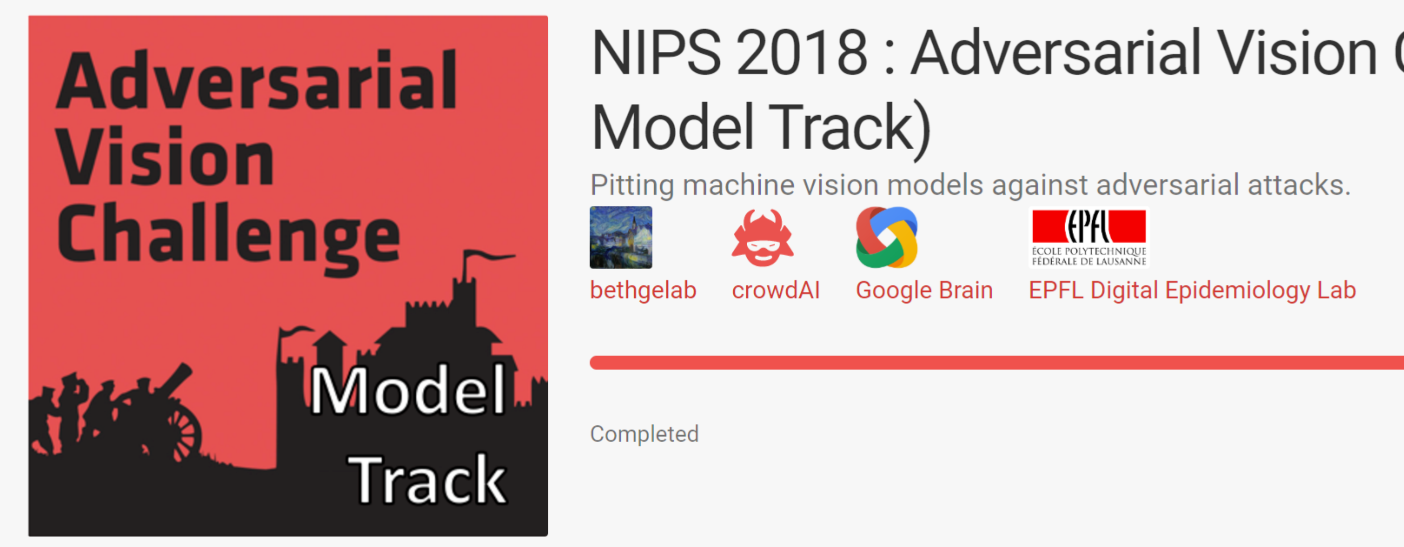
Nicholas Carlini David Wagner

Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples

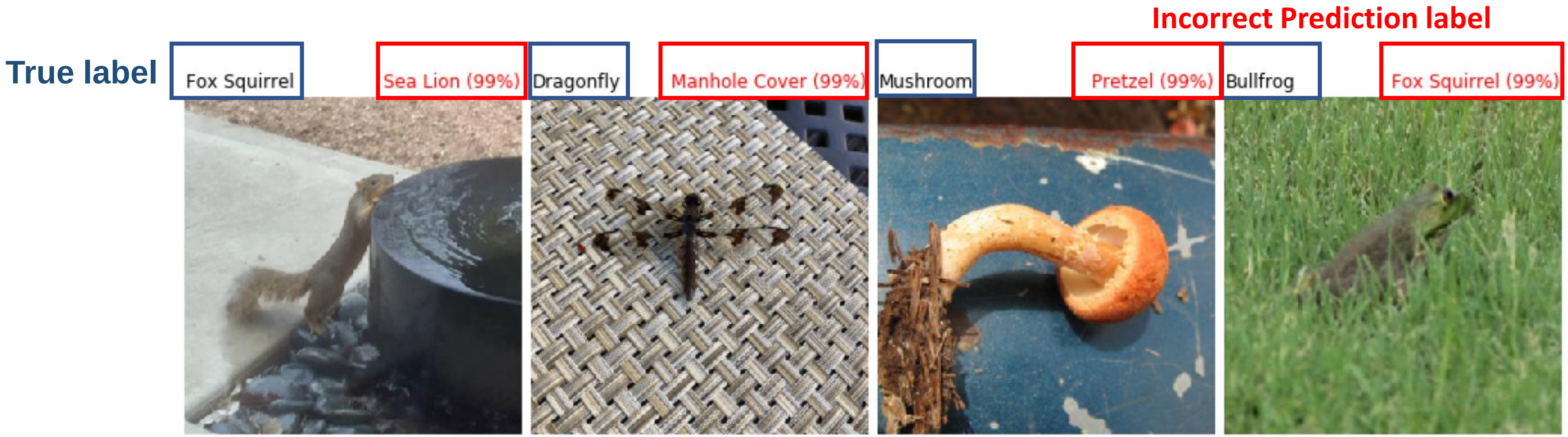
---

Anish Athalye [*1] **Nicholas Carlini** [*2] **David Wagner** [2]

Attack and Defense Arms Race



“Natural Adversarial Examples”



Dan Hendrycks, et al., Natural Adversarial Examples, arXiv, 2019

Where we are and where we go

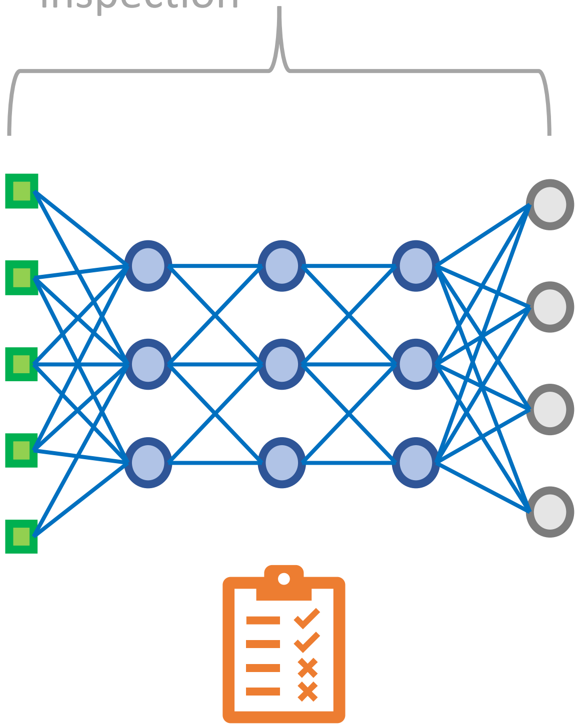
- A defense is robust only when it is known to an adversary but still cannot break it (defender makes the first move and is transparent to an attacker)

1. Data augmentation with adversarial examples: helps but did not solve the problem
2. Standard training to robust training (adversarial training):
   - Minimize _{model parameters} Loss(data, labels, model)
   - Minimize_{model parameters} Maximize_{attack} Loss(manipulated(data), labels, model)
   - Effective, but not scalable, significant drop in test accuracy
3. Input transformation, correction & anomaly detection: many are bypassed by advanced attacks
4. New learning model and training loss: slow progress
5. Model with diversity: model ensembles & model with randomness
6. Domain and task-specific defenses: case-by-case, not automated
7. Combination of all the effective methods: system design

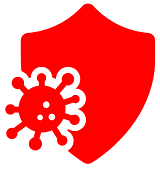
Defenses: Detection and Patching

Trained neural network

- Large models with "good" test performance
- Handful of clean data for inspection

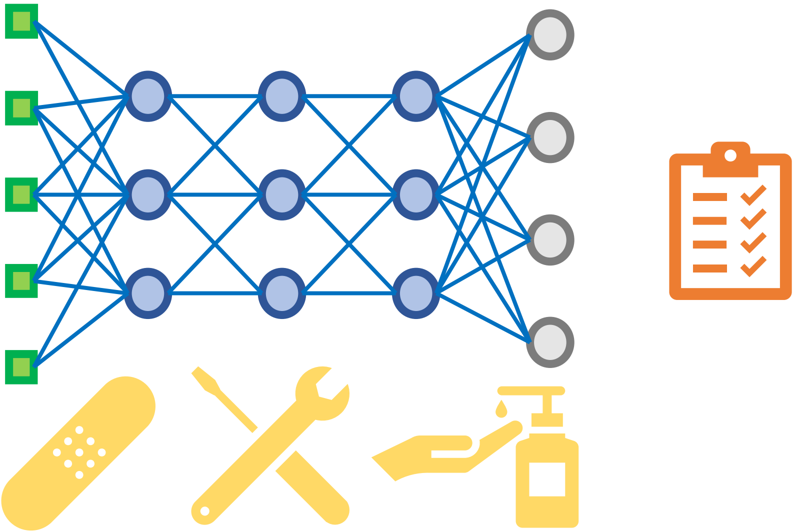


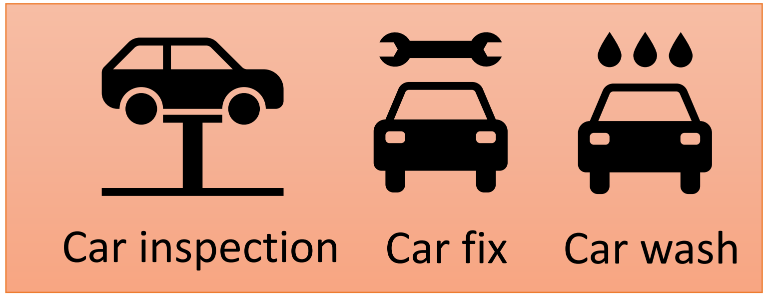
Detection



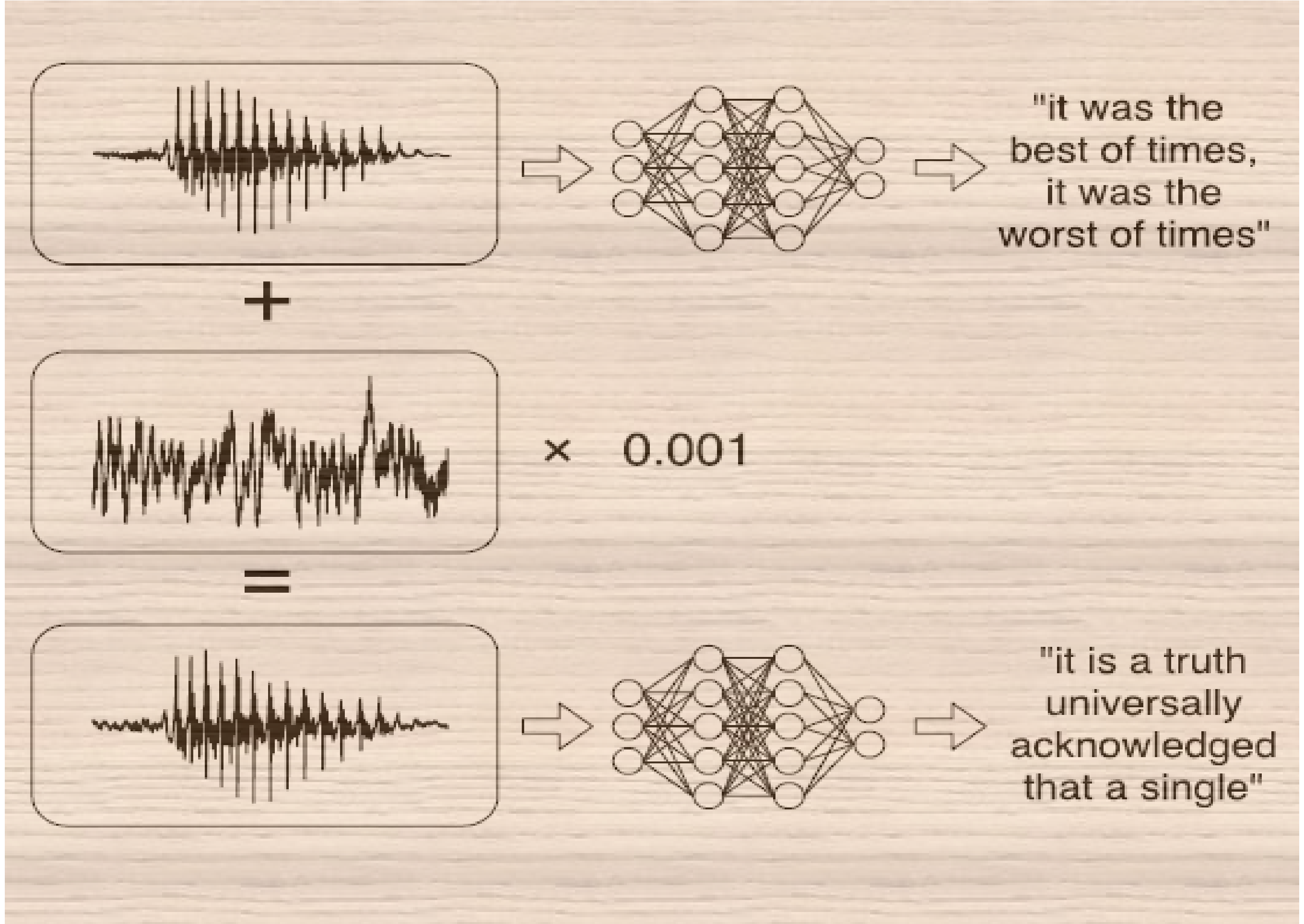
No Trojan found

Patching





Case study: audio adversarial examples



without the dataset the article is useless

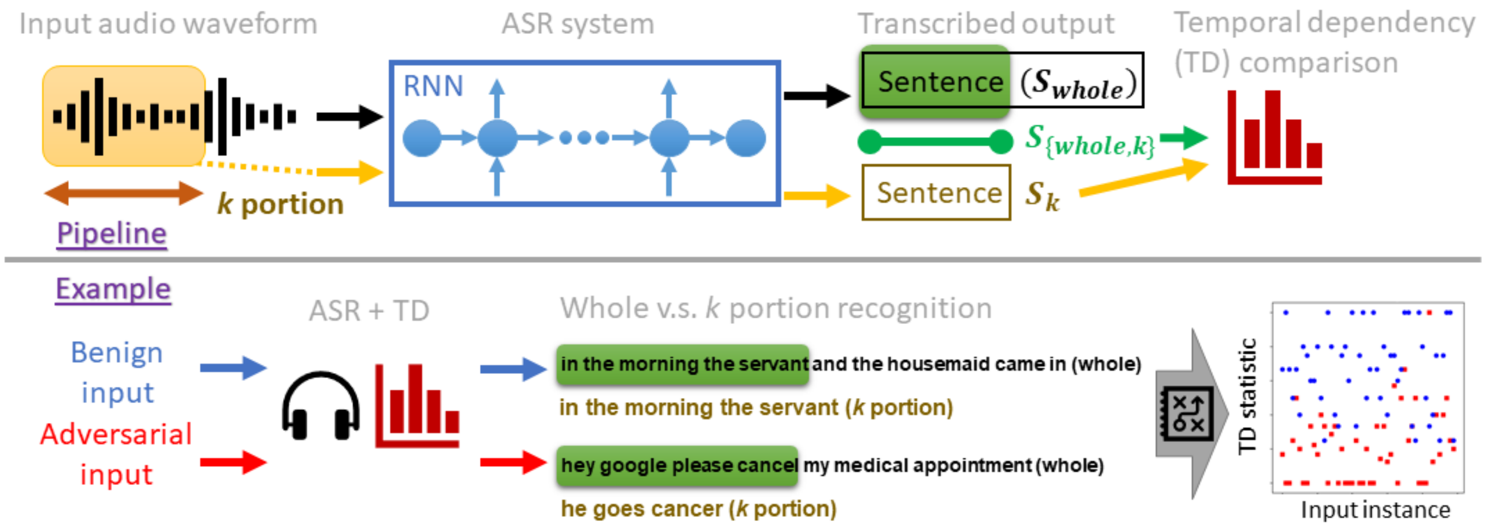
What did your hear?

okay google browse to evil.com

Mitigating audio adversarial attacks

Inference-phase threat

- Leveraging temporal dependency (TD) in audio data to combat audio adversarial examples in automatic speech recognition systems



Type	Transcribed results
Original	then good bye said the rats and they went home
the first half of Original	then good bye said the raps
Adversarial (short)	hey google
First half of Adversarial	he is
Adversarial (medium)	this is an adversarial example
First half of Adversarial	thes on adequate
Adversarial (long)	hey google please cancel my medical appointment
First half of Adversarial	he goes cancer

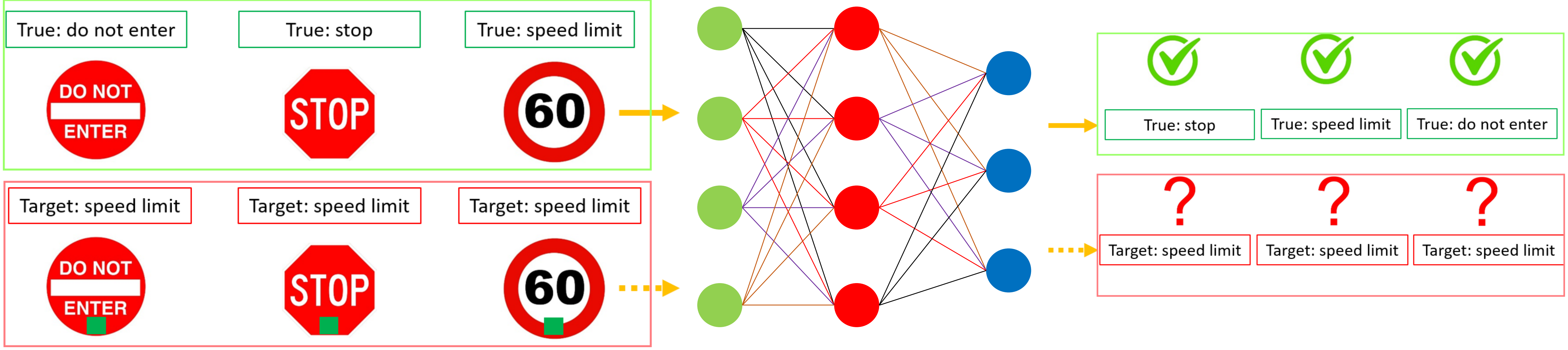
Dataset	LSTM	TD (WER)	TD (CER)	TD (LCP ratio)
Common Voice	0.712	**0.936**	0.916	0.859
LIBRIS	0.645	0.930	**0.933**	0.806

Characterizing Audio Adversarial Examples Using Temporal Dependency. Zhuolin Yang, Bo Li, Pin-Yu Chen and Dawn Song. ICLR 2019

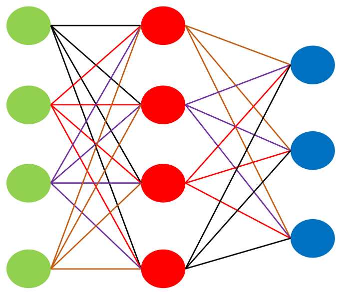
Training-phase threat

Can I know a trained model has Trojan (backdoor)?

Adversary trains a Trojan model using clean data + poisoned data and release the trained model



Trojan trigger



Task: does a given model has backdoor?

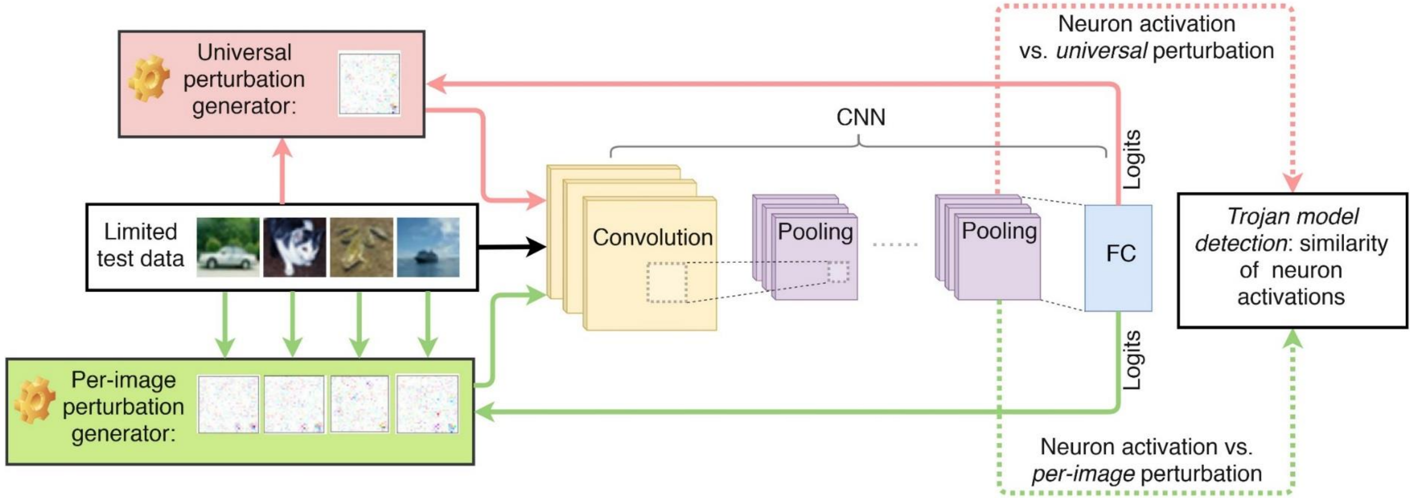
Credit: Ren Wang @ RPI

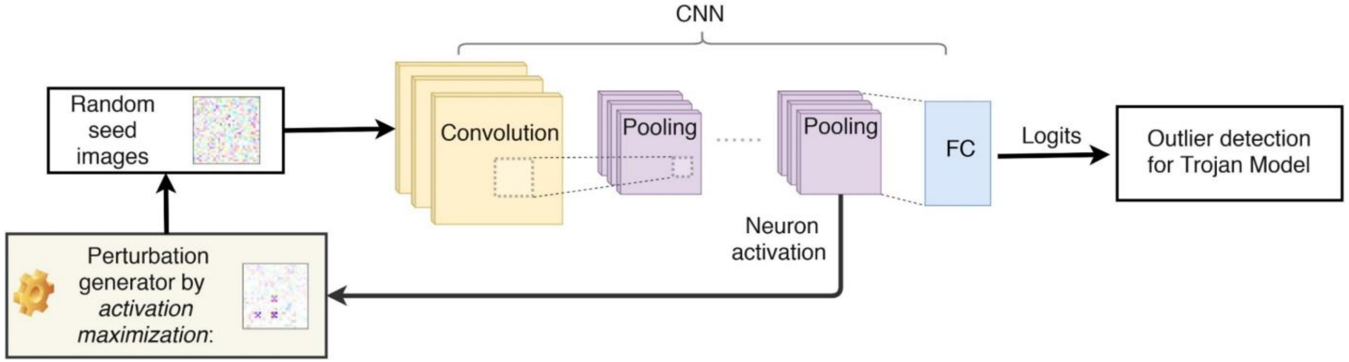
IBM Research AI

Inference-phase threat

Practical Detection of Trojan Models with Limited Data

- Data-limited TrojanNet Detector:
  - only requires one sample per class
  - nearly perfect detection performance
- Data-free TrojanNet Detector:
  - does not require any data
  - uses neural activation maximization
- Shortcut hypothesis: Our detector compares similarity between per-sample perturbation and universal perturbation (shortcut)
- Our detector can generate potential trigger patterns and targeted labels for inspection





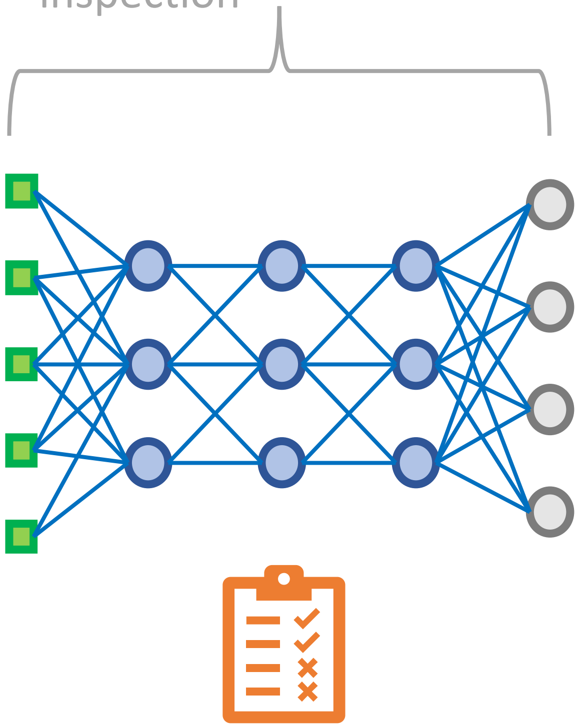
		DL-TND (clean)	DL-TND (Trojan)	NC (clean)	NC (Trojan)
CIFAR-10	ResNet-50	20/20	20/20	11/20	13/20
	VGG16	10/10	9/10	5/10	6/10
	AlexNet	10/10	10/10	6/10	7/10
GTSRB	ResNet-50	12/12	12/12	10/12	6/12
	VGG16	9/9	9/9	6/9	7/9
	AlexNet	9/9	8/9	5/9	5/9
ImageNet	ResNet-50	5/5	5/5	4/5	1/5
	VGG16	5/5	4/5	3/5	2/5
	AlexNet	4/5	5/5	4/5	1/5
Total		**84**/85	**82**/85	54/85	48/85

Ren Wang, Gaoyuan Zhang, Sijia Liu, Pin-Yu Chen, Jinjun Xiong, and Meng Wang. Practical Detection of Trojan Neural Networks: Data-Limited and Data-Free. *ECCV 2020*

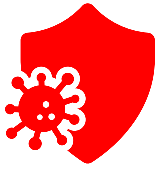
Defenses: Detection and Patching

Trained neural network

- Large models with "good" test performance
- Handful of clean data for inspection

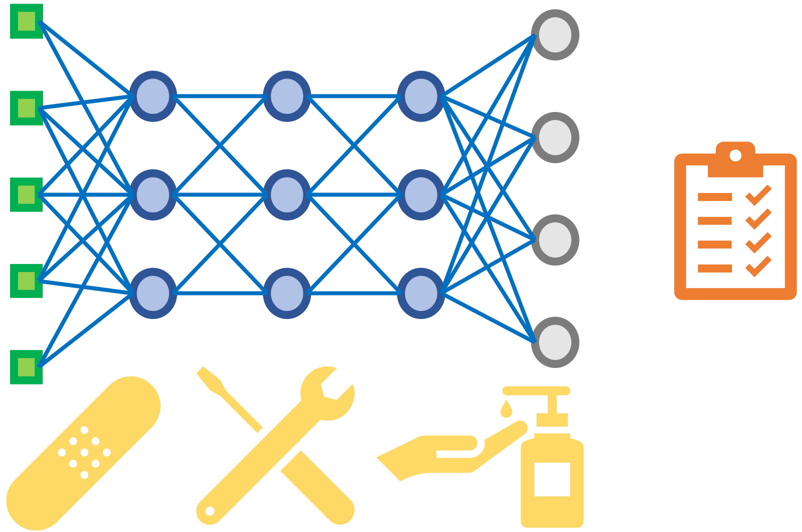


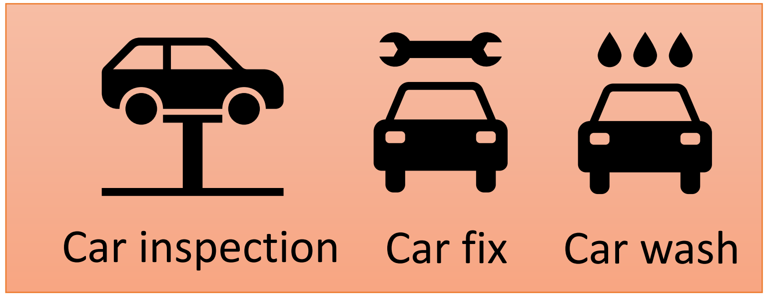
Detection



No Trojan found

Patching



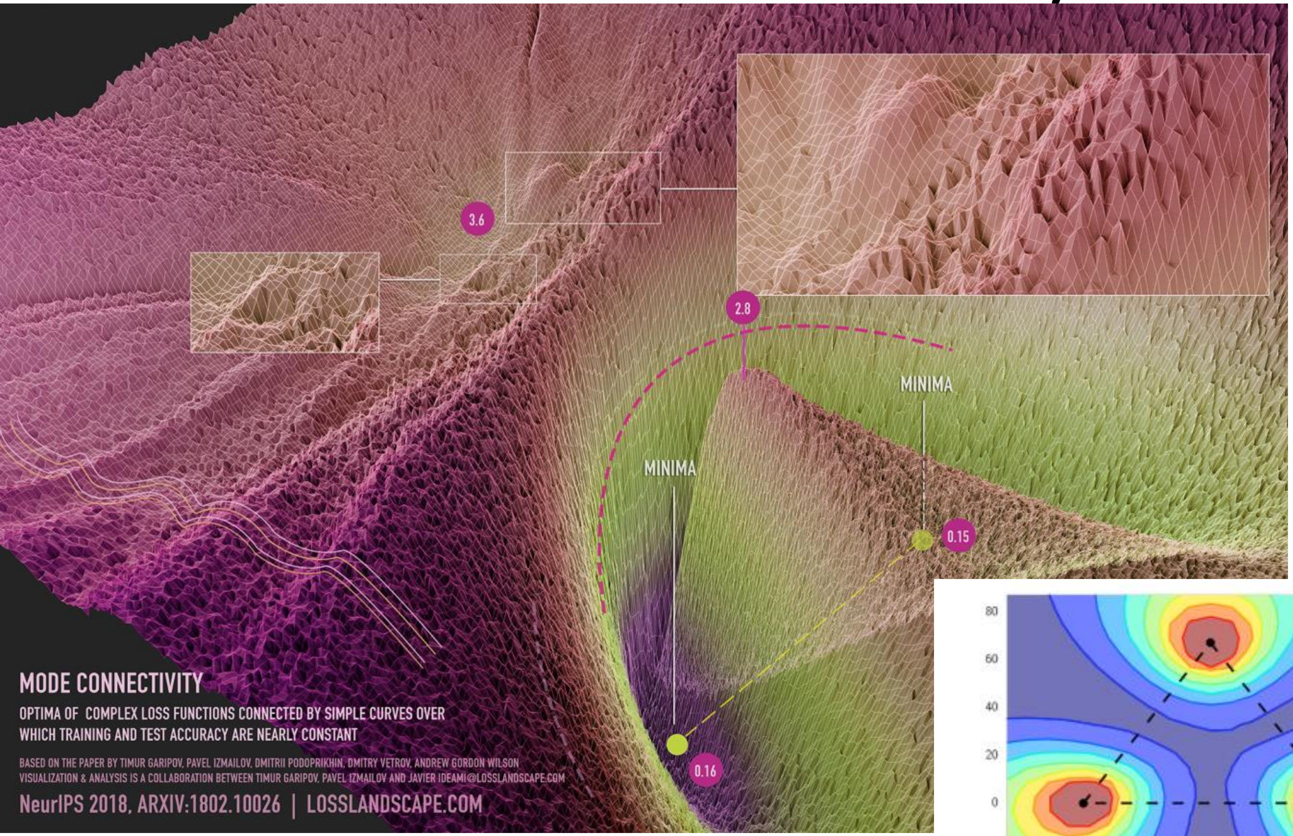


Problem Setup: Trusted Finetuning with Limited Data

Training-phase threat

- Given a model from an untrusted source, can one use a small set of clean and trusted data samples to sanitize the model, in order to alleviate the potential backdoor effect while maintaining similar performance on regular task?
- The size of trusted data samples should be limited, otherwise training from scratch outweighs the risk of using tampered models
- This problem is beyond detecting backdoor models (post-detection phase) -> Model recovery instead of model detection

Mode Connectivity in Loss Landscape



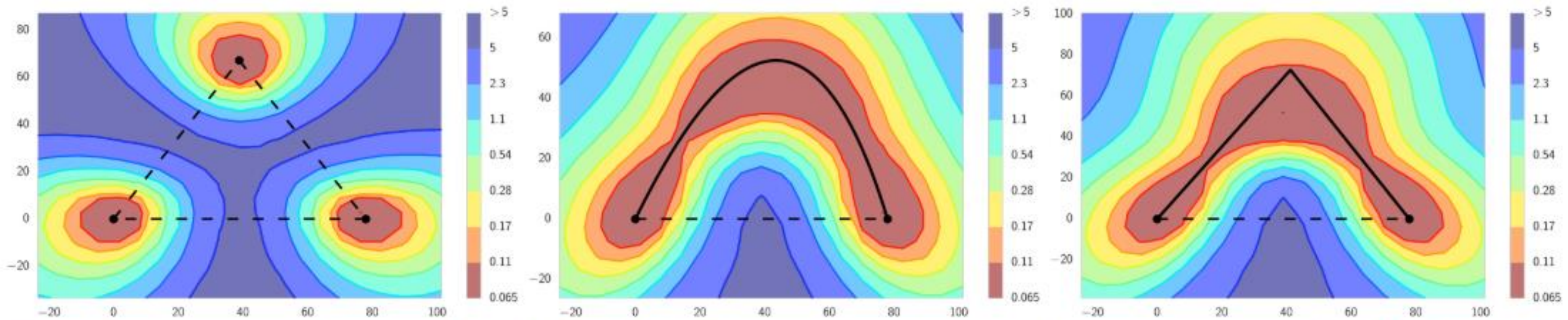


Figure 2: Loss surface of ResNet-164 on CIFAR-100. **Left**: three optima for independently trained networks; **Middle** and **Right**: A quadratic Bezier curve, and a polygonal chain with one bend, connecting the lower two optima on the left panel along a path of near-constant loss.

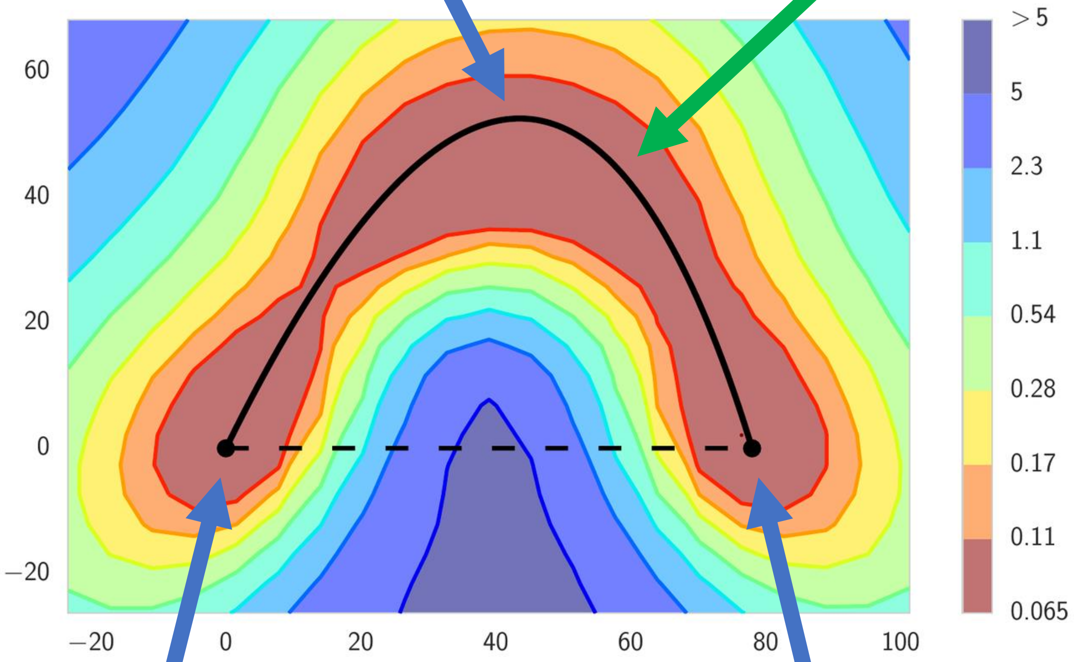
https://izmailovpavel.github.io/curves_blogpost/

Trusted Finetuning / Model Sanitization

- Quadratic Bezier Curve:
$$\phi_\theta(t) = (1-t)^2\omega_1 + 2t(1-t)\theta + t^2\omega_2$$
$$0 \le t \le 1$$
- Training loss:
$$L(\theta) = \mathrm{E}_{t\sim Unif[0,1]} loss(\phi_\theta(t))$$
- Use stochastic optimization on the trusted dataset to update θ
- How do we start with two trained models? (see paper)
- Neuron alignment improves mode connectivity

Parametrized simple curve function $\phi_\theta(t), 0 \le t \le 1$

Sanitized models



Trained Model 1 ($t = 0$); denoted by ω_1

Trained Model 2 ($t = 1$); denoted by ω_2

Pu Zhao, Pin-Yu Chen, Payel Das, Karthikeyan Natesan Ramamurthy, and Xue Lin. Bridging Mode Connectivity in Loss Landscapes and Adversarial Robustness. ICLR 2020

N. Joseph Tatro, Pin-Yu Chen, Payel Das, Igor Melnyk, Prasanna Sattigeri, and Rongjie Lai. Optimizing Mode Connectivity via Neuron Alignment. NeurIPS 2020

Mode Connectivity Provides Good Prior for Trusted Finetuning with few clean data

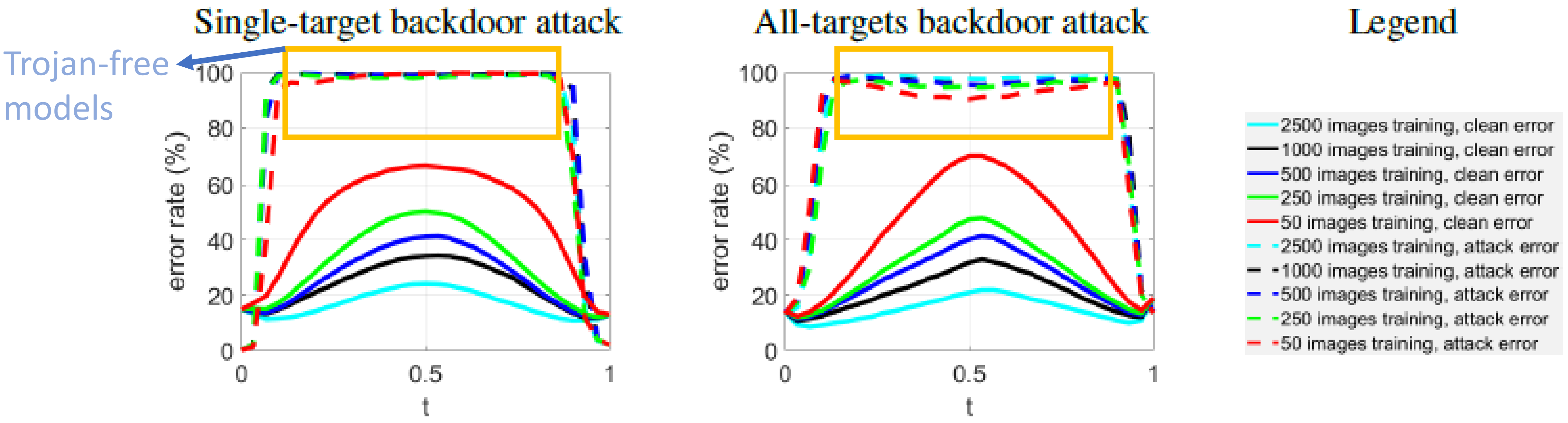


Figure 2: Error rate against backdoor attacks on the connection path for CIFAR-10 (VGG). The error rate of clean/backdoored samples means the standard-test-error/attack-failure-rate, respectively.

Trusted Finetuning Outperforms Baselines

- Baselines: (i) Finetuning (ii) Train from scratch (iii) Weight Pruning+Finetuning (iv) random Gaussian perturbation to model weights
  - ❑ Train from Scratch removes backdoor but has low clean accuracy
  - ❑ Pruning remains high clean accuracy but suffers high attack success rate
  - ❑ Finetuning is suboptimal when the data size is limited

Table 2: Performance against single-target backdoor attack. The clean/backdoor accuracy means standard-test-accuracy/attack-success-rate, respectively. More results are given in Appendix E.

		Method / Bonafide data size	2500	1000	500	250	50
CIFAR-10 (VGG)	Clean Accuracy (Higher is better)	Path connection ($t = 0.1$)	88%	83%	80%	77%	63%
		Fine-tune	84%	82%	78%	74%	46%
		Train from scratch	50%	39%	31%	30%	20%
		Noisy model ($t = 0$)	21%	21%	21%	21%	21%
		Noisy model ($t = 1$)	24%	24%	24%	24%	24%
		Prune	88%	85%	83%	82%	81%
	Backdoor Accuracy (Lower is better)	Path connection ($t = 0.1$)	1.1%	0.8%	1.5%	3.3%	2.5%
		Fine-tune	1.5%	0.9%	0.5%	1.9%	2.8%
		Train from scratch	0.4%	0.7%	0.3%	3.2%	2.1%
		Noisy model ($t = 0$)	97%	97%	97%	97%	97%
		Noisy model ($t = 1$)	91%	91%	91%	91%	91%
		Prune	43%	49%	81%	79%	82%

- ✓ Ours maintains superior accuracy on clean data while simultaneously attaining low attack accuracy
- ✓ The success of using mode connectivity is NOT by chance: 1000 noisy models suffer from low clean accuracy and high attack success rate

Inference-phase threat

Adversarial Training and Benchmarks

Towards Deep Learning Models Resistant to Adversarial Attacks

Aleksander Mądry* MIT madry@mit.edu
Aleksandar Makelov* MIT amakelov@mit.edu
Ludwig Schmidt* MIT ludwigs@mit.edu
Dimitris Tsipras* MIT tsipras@mit.edu
Adrian Vladu* MIT avladu@mit.edu

ICLR'18

Theoretically Principled Trade-off between Robustness and Accuracy

Hongyang Zhang* CMU & TTIC hongyanz@cs.cmu.edu
Yaodong Yu† University of Virginia yy8ms@virginia.edu
Jiantao Jiao UC Berkeley jiantao@eecs.berkeley.edu
Eric P. Xing CMU & Petuum Inc. epxing@cs.cmu.edu
Laurent El Ghaoui UC Berkeley elghaoui@berkeley.edu
Michael I. Jordan UC Berkeley jordan@cs.berkeley.edu

ICML'18

RobustBench Leaderboards Paper FAQ Contribute Model Zoo

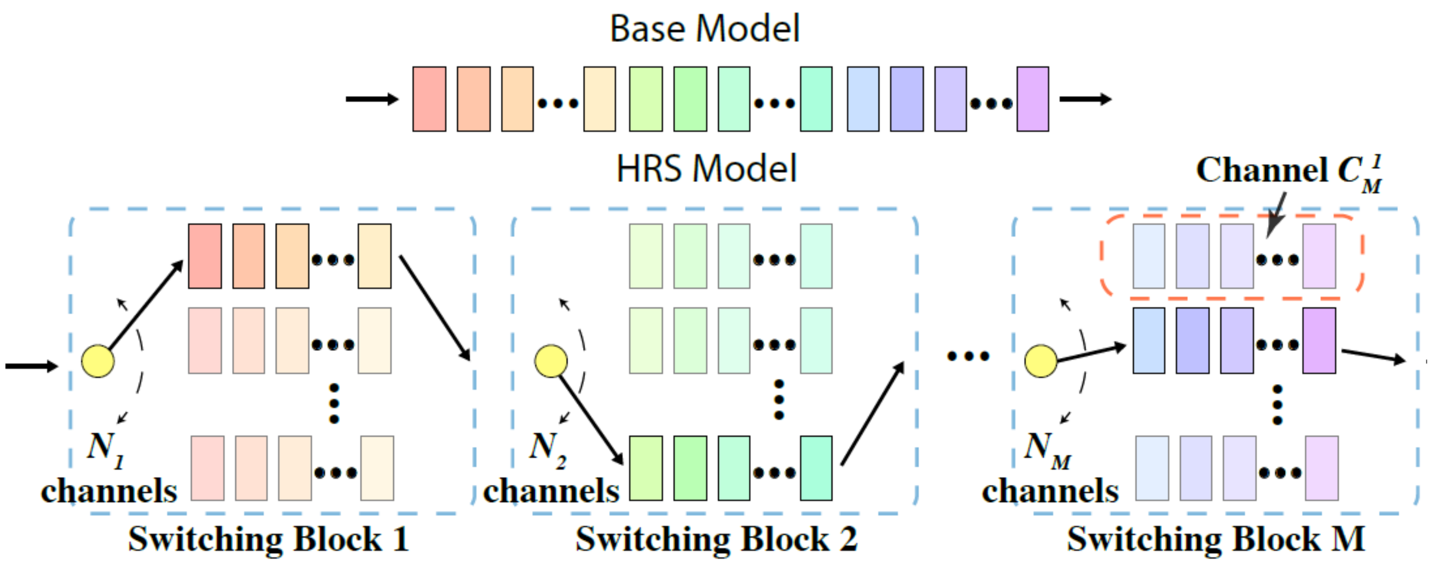
RobustBench

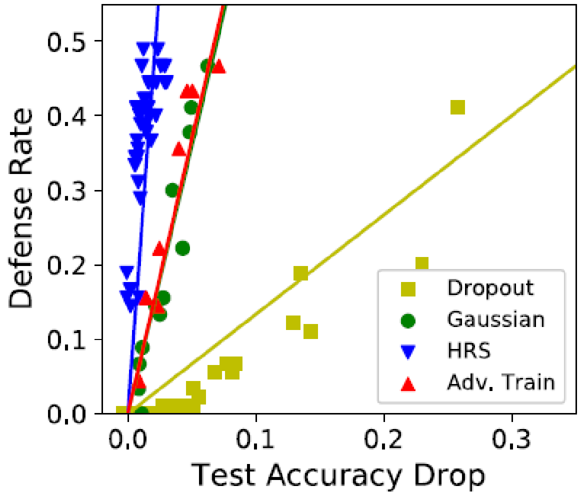
A standardized benchmark for adversarial robustness

- Adversarial training: $min_{\theta} \sum_{i=1}^{n} max_{\{\delta_i\}_{i=1}^{n}, ||\delta_i|| \le \epsilon} loss(x_i + \delta_i, y_i \,;\, \theta)$
- TRADES: $min_{\theta} \sum_{\{i=1\}}^{n} loss(x_i + \delta_i, y_i \,; \theta) + \lambda \cdot max_{\{\delta_i\}_{i=1}^{n}, ||\delta_i|| \le \epsilon} loss(f_\theta(x_i), f_\theta(x_i + \delta_i); \theta)$
- Use of unlabeled data or pretraining can improve adversarial robustness
- Adaptive attack and Auto attack; RobustBench

HRS Training: Hierarchical Random Switching

- A randomness-driven training method that achieves 5X better robustness-accuracy trade-off than SOTA





Defense Methods	Mean DES
Dropout	0.815
Gaussian	6.176
Adv. Train	7.367
HRS	35.429

Xiao Wang*, Siyue Wang*, Pin-Yu Chen, Yanzhi Wang, Brian Kulis, Xue Lin, and Sang Chin, "Protecting Neural Networks with Hierarchical Random Switching: Towards Better Robustness-Accuracy Trade-off for Stochastic Defenses," IJCAI 2019

Inference-phase threat

SPROUT: Self-Progressing Robust Training

Minhao Cheng, Pin-Yu Chen, Sijia Liu, Shiyu Chang, Cho-Jui Hsieh, Payel Das. AAAI 2021

CAT: Customized Robust Training for Improved Robustness

Minhao Cheng, Qi Lei, Pin-Yu Chen, Inderjit Dhillon, Cho-Jui Hsieh

SPROUT: Self-Progressing Robust Training

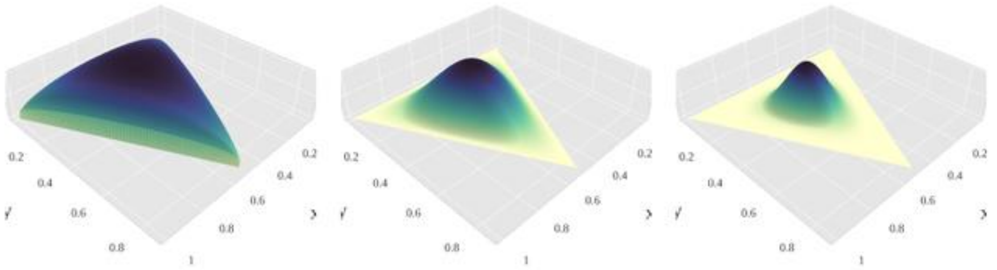
- Observation: static label smoothing during training improves adversarial robustness
- Label smoothing: instead of model training on one-hot coded labeled data samples $\{x_i, y_i\}_{i=1}^{n}$, we train on $\{x_i, \tilde{y}_i\}_{i=1}^{n}$, where

$$\tilde{y} = (1 - \alpha)y + \alpha \cdot u\,, \quad \alpha \in (0,1)$$

- In practice, $u = \frac{1}{K}\mathbf{1}$ (i.e. uniform label smoothing)
- Pros: Attack-independent training, efficient
- Cons: Marginal robustness gain compared to adversarial training

Self-Progressing Robust Training. Minhao Cheng, Pin-Yu Chen, Sijia Liu, Shiyu Chang, Cho-Jui Hsieh, Payel Das. AAAI 2021

Dirichlet Label Smoothing



- Our proposed parameterized label technique
- Draw training label from a parameterized distribution:

$$\tilde{y} = (1-\alpha)y + \alpha \cdot Dirichlet(\beta)$$

- Self-progressing training with Dirichlet label smoothing:

$$min_\theta max_\beta \sum_{i=1}^{n} loss(x_i, \tilde{y}_i\ ;\ \theta, \beta)$$

- Recall Adversarial Training [Madry ICLR'18]:

$$min_\theta \sum_{i=1}^{n} max_{\{\delta_i\}_{i=1}^{n}} loss(x_i + \delta_i, y_i\ ;\ \theta)$$

SPROUT = Dirichlet LS + Gaussian Augmentation + Mixup - Attack Independent!

- Dirichlet LS: $\tilde{y} = (1-\alpha)y + \alpha \cdot Dirichlet(\beta)$
- Gaussian Augmentation: $\tilde{x} = x + \mathrm{N}(0, \sigma^2 \mathrm{I})$
- Mixup of two data samples $\{x_i, y_i\}, \{x_j, y_j\}$:

$$\tilde{x} = \lambda x_i + (1-\lambda)x_j, \tilde{y} = \lambda y_i + (1-\lambda)y_j \quad , \quad \lambda \in (0,1)$$

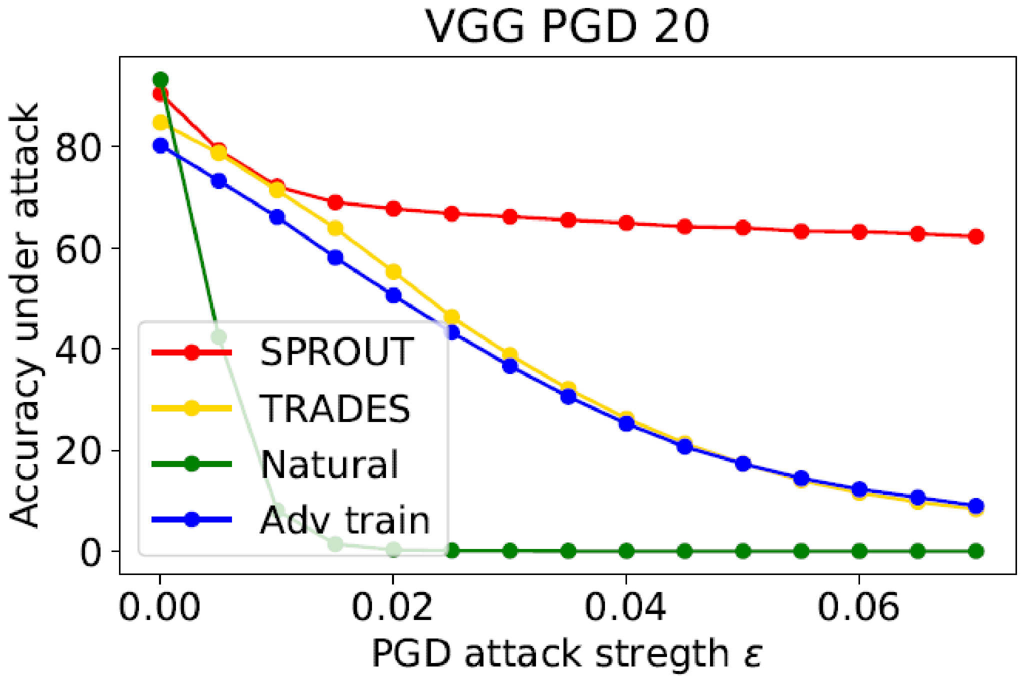
- Overall training objective: $min_\theta max_\beta \sum_{i=1}^{n} loss(\tilde{x}_i, \tilde{y}_i \ ; \ \theta, \beta | x_i, y_i)$
- These three techniques are free of attack-generation
- We will show the robustness gains from these three methods are complimentary

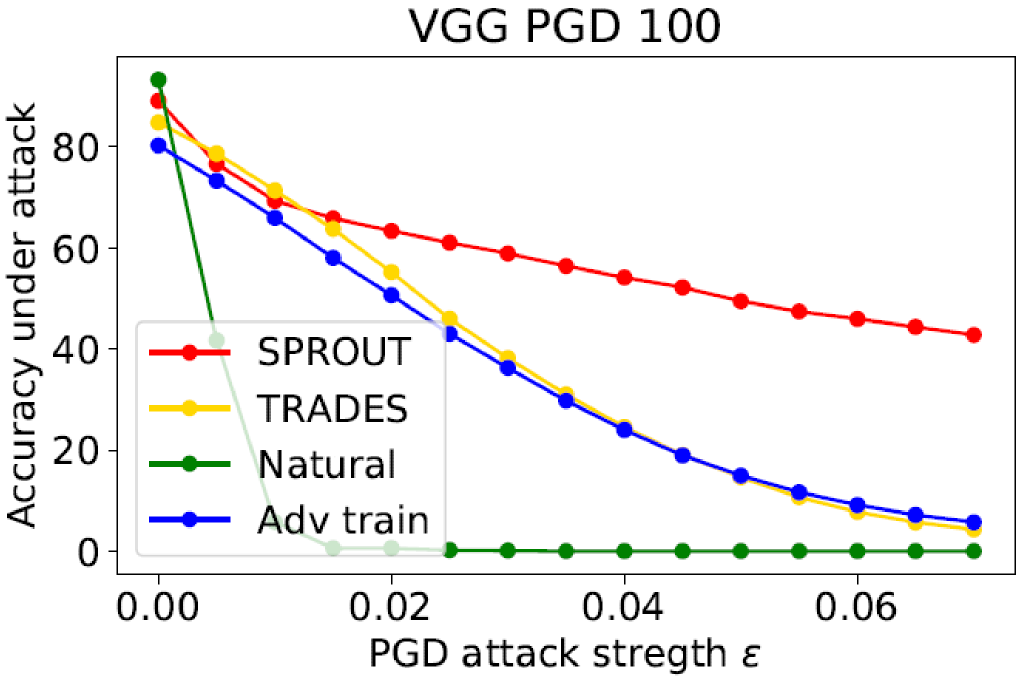
mixup: Beyond Empirical Risk Minimization. Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, David Lopez-Paz. ICLR 2018

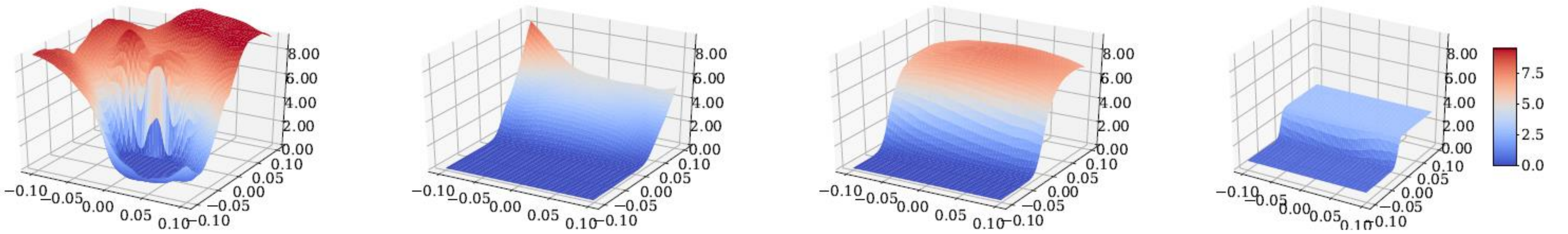
Algorithm 1 SPROUT algorithm

Input: Training dataset (X, Y), Mixup parameter λ, Gaussian augmentation variance Δ^2, model learning rate γ_θ, Dirichlet label smoothing learning rate $\gamma_{\boldsymbol{\beta}}$ and parameter α, cross entropy loss L
Initial model θ: random initialization (train from scratch) or pre-trained model checkpoint
Initial $\boldsymbol{\beta}$: random initialization
for epoch$=1, \ldots, N$ **do**
  for minibatch $X_B \subset X, Y_B \subset Y$ **do**
    $X_B \leftarrow \mathcal{N}(X_B, \Delta^2)$
    $X_{mix}, Y_{mix} \leftarrow \text{Mixup}(X_B, Y_B, \lambda)$
    $Y_{mix} \leftarrow \text{Dirichlet}(\alpha Y_{mix} + (1-\alpha)\boldsymbol{\beta})$
    $g_\theta \leftarrow \nabla_\theta L(X_{mix}, Y_{mix}, \theta)$
    $g_{\boldsymbol{\beta}} \leftarrow \nabla_{\boldsymbol{\beta}} L(X_{mix}, Y_{mix}, \theta)$
    $\theta \leftarrow \theta - \gamma_\theta g_\theta$
    $\boldsymbol{\beta} \leftarrow \boldsymbol{\beta} + \gamma_{\boldsymbol{\beta}} g_{\boldsymbol{\beta}}$
  end for
end for
return θ

Substantial Robustness Improvement

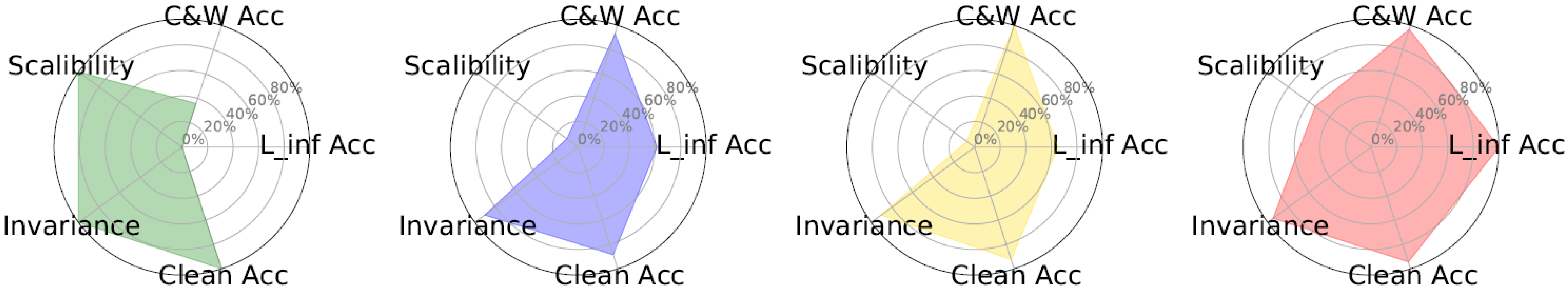






(a) Natural (b) Adv train (c) TRADES (d) SPROUT

Better Scalability and Comprehensive Performance



(a) Natural (b) Adversarial training (c) TRADES (d) SPROUT (ours)

Customized Adversarial Training (CAT)

- Recall Adversarial Training [Madry ICLR'18]:

$$min_\theta \sum_{i=1}^{n} max_{\{\delta_i\}_{i=1}^{n},\, \|\delta_i\| \leq \epsilon} loss(x_i + \delta_i, y_i \,;\, \theta)$$

- Not all samples should be treated equally in adversarial training
- Nor all their training labels
- Our CAT formulation:

$$min_\theta \sum_{i=1}^{n} max_{\{\delta_i\}_{i=1}^{n},\, \|\delta_i\| \leq \epsilon_i} loss(x_i + \delta_i, \tilde{y}_i \,;\, \theta)$$

How does CAT work? Self-Progressing!

- $min_{\theta} \sum_{i=1}^{n} max_{\{\delta_i\}_{i=1}^{n}, \|\delta_i\| \leq \epsilon_i} loss(x_i + \delta_i, \tilde{y}_i \; ; \; \theta)$

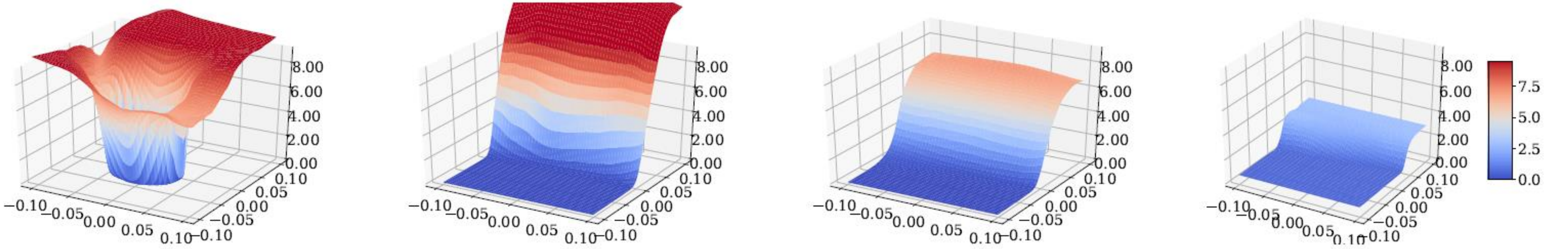
$$\tilde{y}_i = (1 - c\epsilon_i)y_i + c\epsilon_i \mathrm{Dirichlet}(1)$$

The model prediction should be less confident for perturbed samples $x_i + \delta_i$ that are further away from x_i

1. Initialize ϵ_i with $\epsilon_i = 0$
2. In each epoch, if $x_i + \delta_i$ still can be classified correctly as y_i , increase ϵ_i (to a maximum value) , otherwise decrease
3. Assign training label $\tilde{y}_i = (1 - c\epsilon_i)y_i + c\epsilon_i \mathrm{Dirichlet}(1)$ to $x_i + \delta_i$
4. Update model θ with $\{x_i + \delta_i, \tilde{y}_i\}$
5. Repeat 2 to 4

CIFAR-10 results

Methods	Clean accuracy	PGD accuracy	C&W accuracy
Natural training	**95.93%**	0%	0%
Adversarial training (Madry et al., 2018)	87.30%	52.68%	50.73%
Dynamic adversarial training (Wang et al., 2019)	84.51%	55.03%	51.98%
TRADES (Zhang et al., 2019b)	84.22%	56.40%$^{(20)}$	51.98%
Bilateral Adv Training (Wang, 2019)	91.00%	57.5%$^{(*20)}$	56.2%$^{(*20)}$
MMA (Ding et al., 2018)	84.36%	47.18%	✗
MART (Wang, 2020)	84.17%	58.56%$^{(20)}$	54.58%
IAAT (Balaji et al., 2019)	91.34%	48.53%$^{(*10)}$	56.80%
CAT-CE (ours)	93.48%	**73.38%**$^{(*20)}$	61.88%$^{(*20)}$
CAT-MIX (ours)	89.61%	73.16%$^{(*20)}$	**71.67%**$^{(*20)}$



(a) Natural (b) Adv train (c) TRADES (d) CAT

Robustness Certification and Evaluation

Certificate for a data sample: For a given model θ and a given data sample x, provide a certificate ϵ for a threat model (e.g. norm-based perturbation $||\delta||$) such that the model prediction of the data sample will not be altered as long as the attack strength is no greater than ϵ : $\boldsymbol{pred(x|\theta) = pred(x+\delta|\theta)}$ **for any** $\boldsymbol{||\delta|| \leq \epsilon}$

How do we evaluate adversarial robustness?

- Game-based approach

❑Specify a set of players (attacks and defenses)

❑Benchmark the performance against each attacker-defender pair

- The metric/rank could be exploited;

No guarantee on unseen threats and future attacks

- Verification-based approach

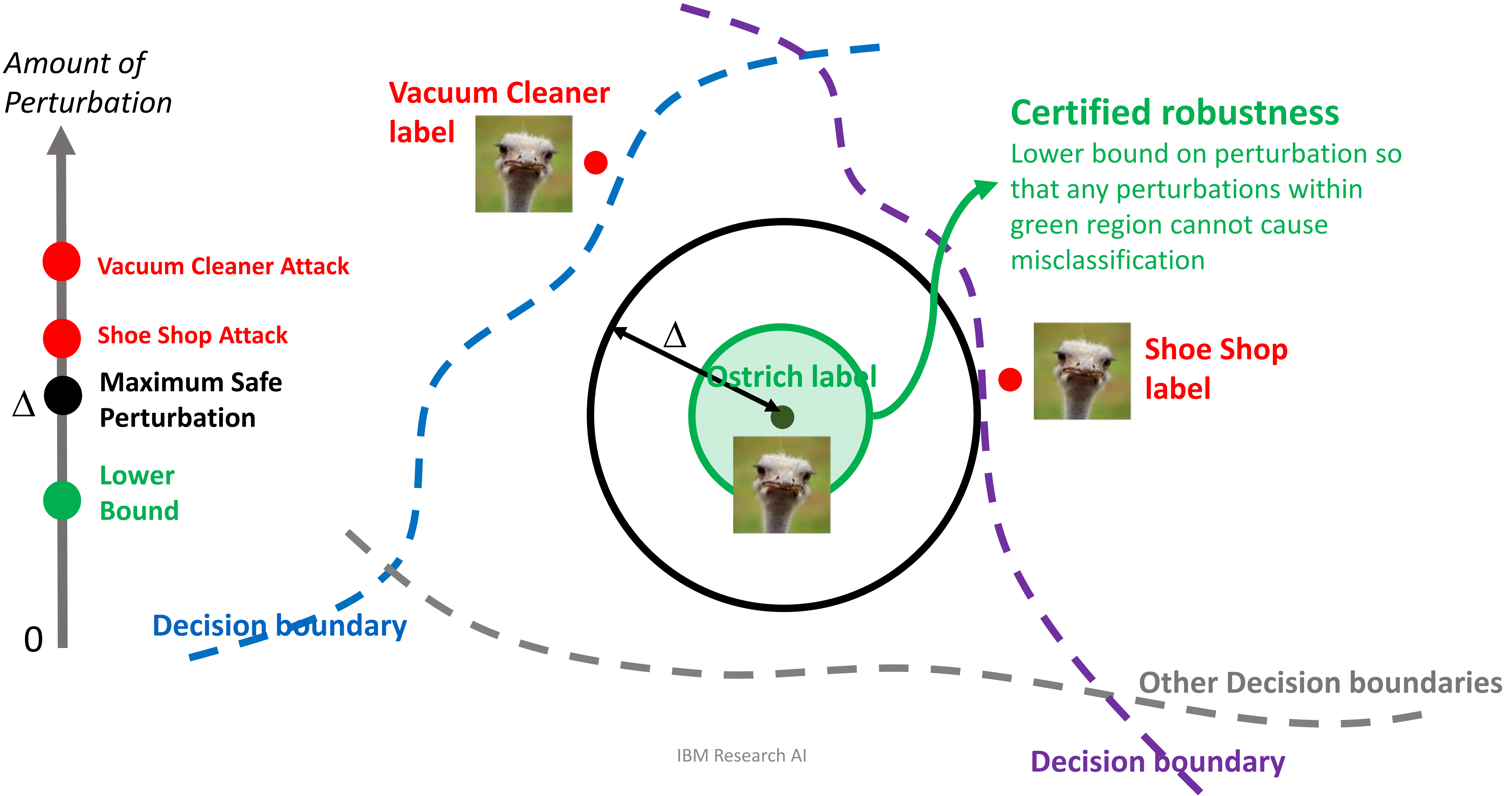
❑Attack-independent: does not use attacks for evaluation

❑Can provide a robustness certificate for safety-critical or reliability-sensitive applications: e.g., no attacks can alter the decision of the AI model if the attack strength is limited

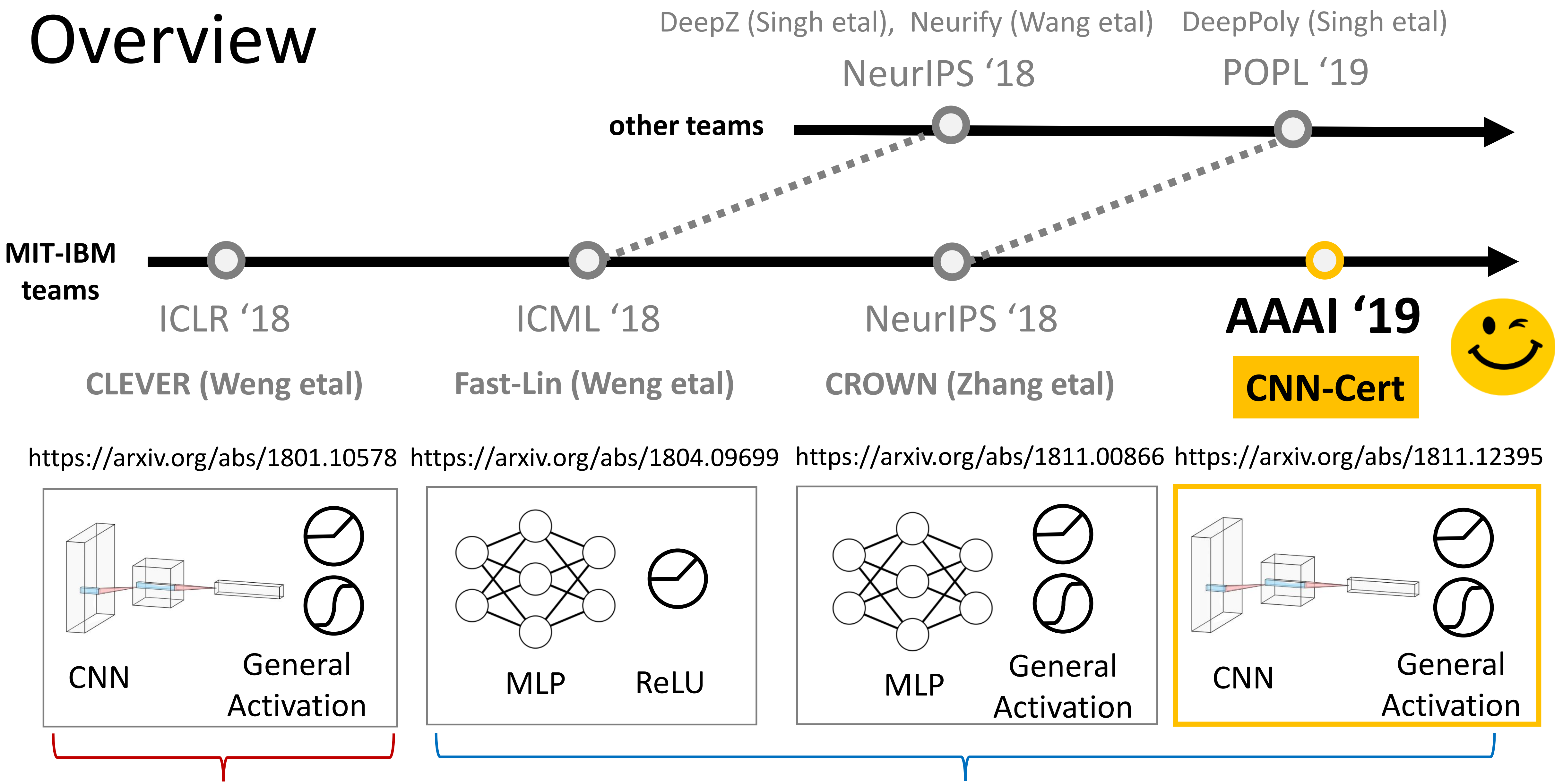
Optimal verification is provably difficult for large neural nets – computationally impractical

- Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks, Guy Katz, Clark Barrett, David Dill, Kyle Julian, Mykel Kochenderfer, CAV 2017
- Efficient Neural Network Robustness Certification with General Activation Functions, Huan Zhang*, Tsui-Wei Weng*, Pin-Yu Chen, Cho-Jui Hsieh, and Luca Daniel, NIPS 2018

Verification: lower bounds on robustness



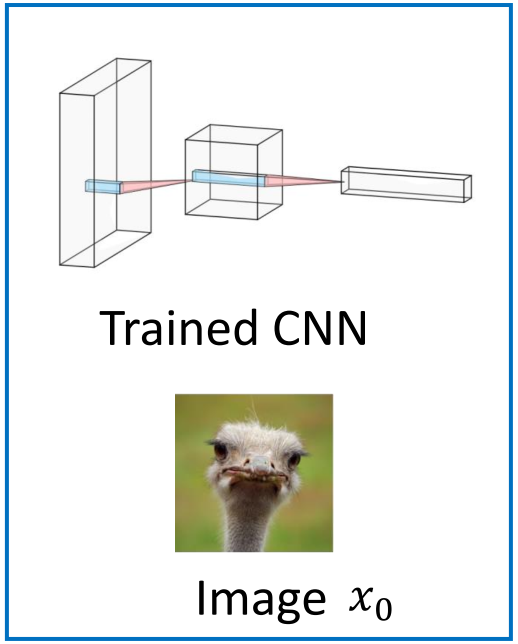
Overview



Robustness Estimation

Robustness Certification

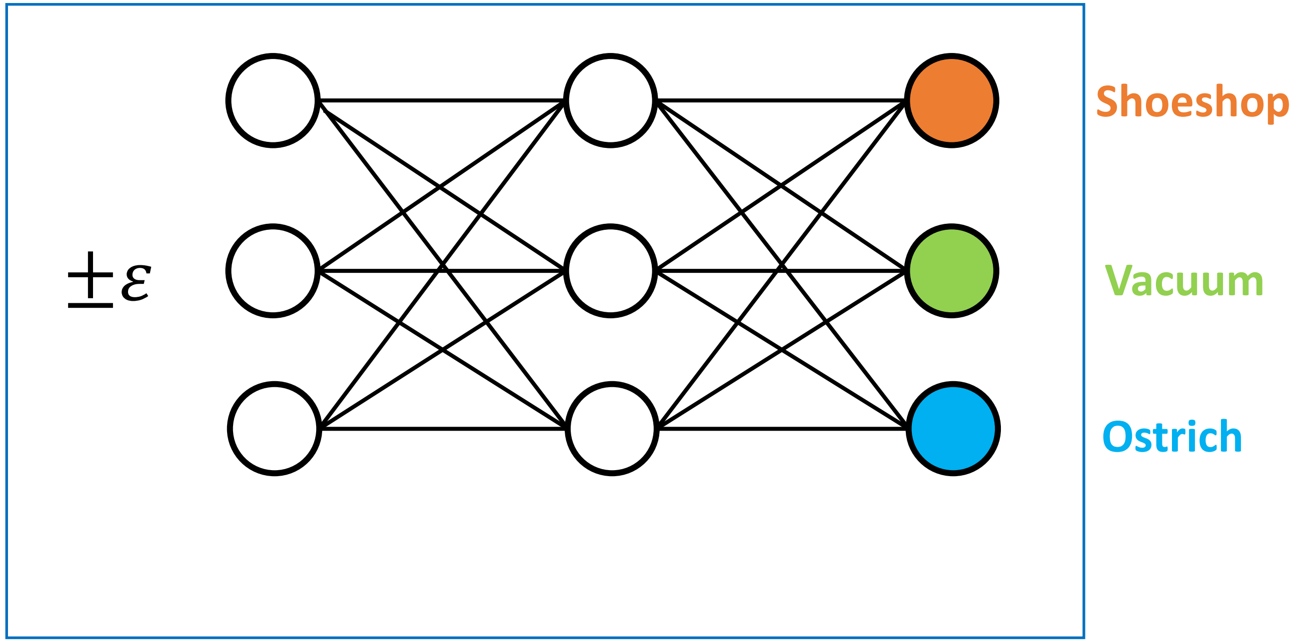
Efficient certified bound with activation bounds



Input

$\|x - x_0\| \leq \varepsilon$

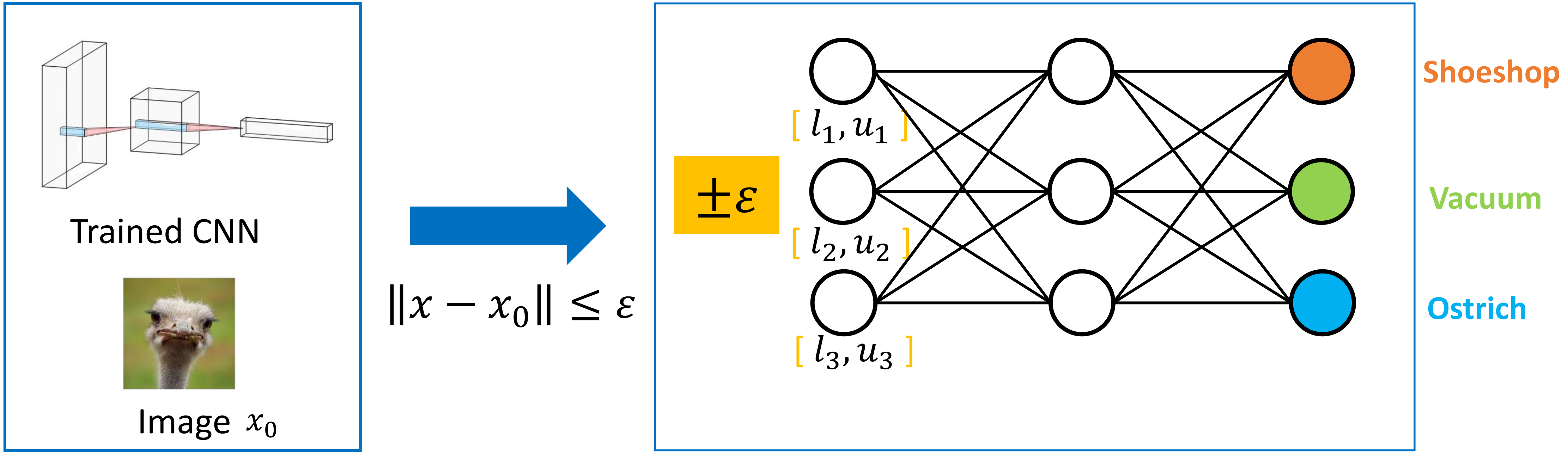
Threat model



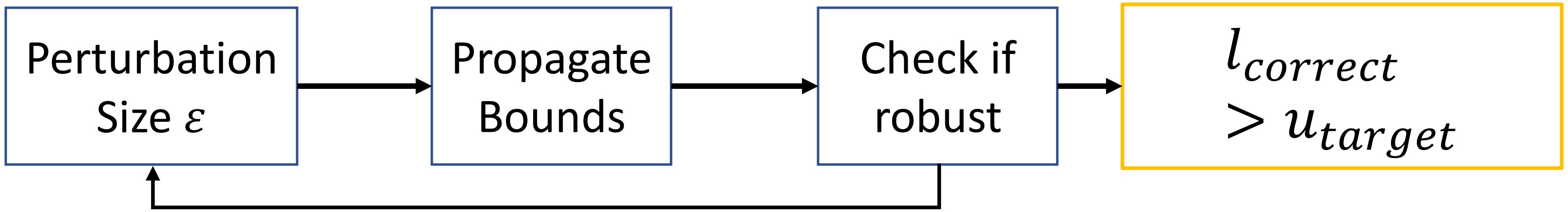
Perturbation Size ε → Propagate Bounds → Check if robust → $l_{correct} > u_{target}$

- Robustness Certificate: Given a data input and a neural network model, under the specified threat model (e.g. L_p norm ball) the top-1 prediction of the perturbed input will not be altered if the perturbation is smaller than $\varepsilon_{certified}$

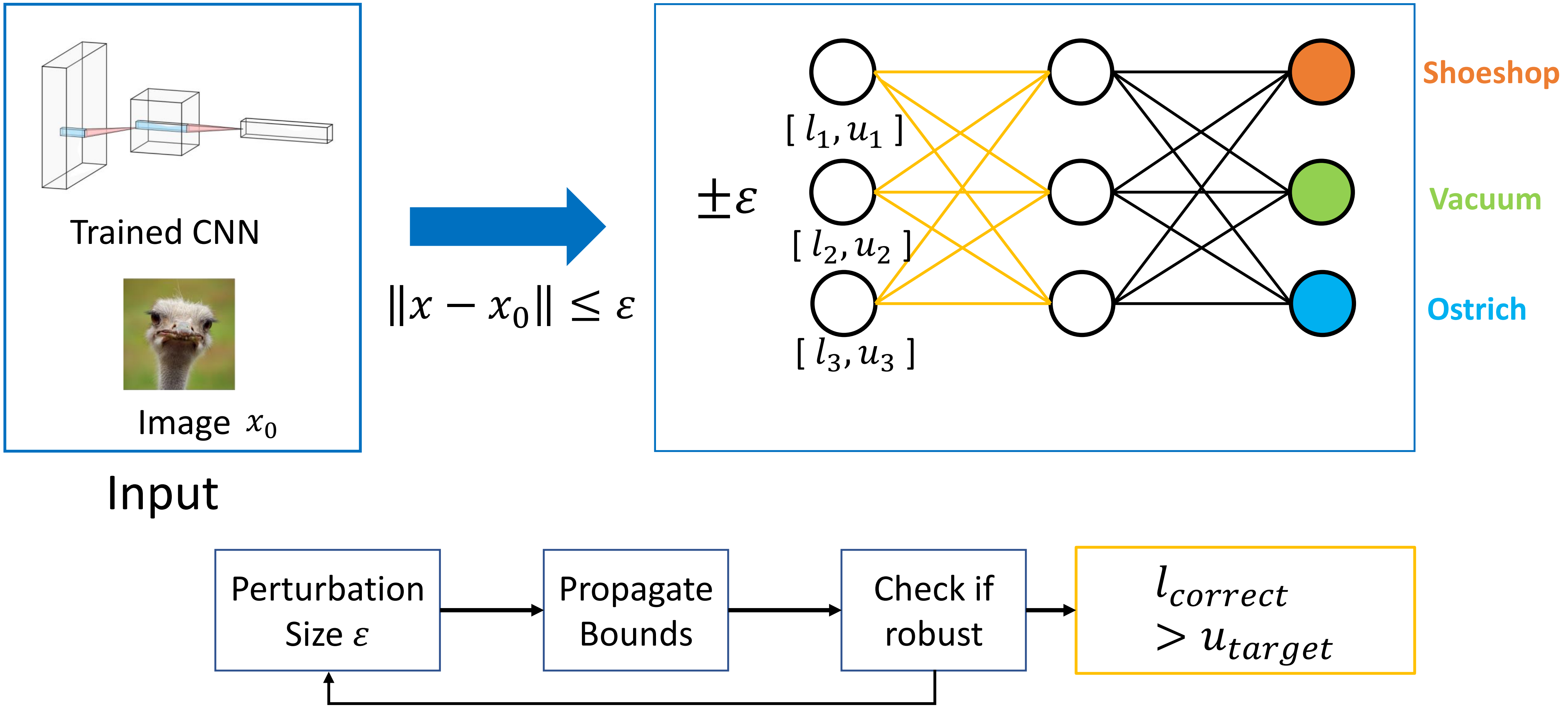
Efficient certified bound with activation bounds



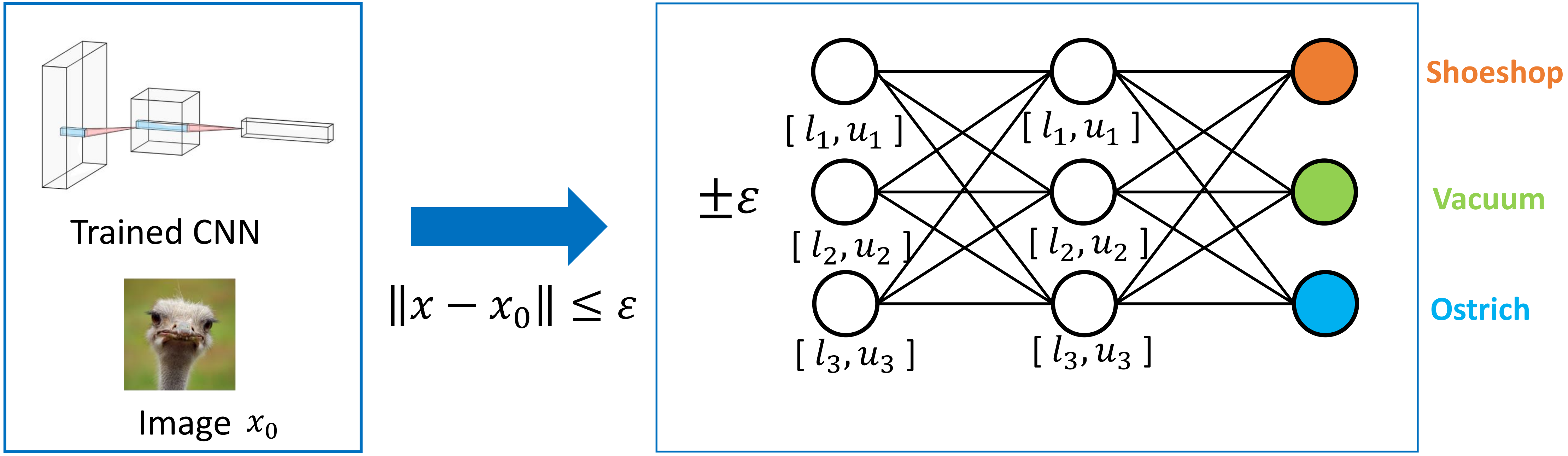
Input



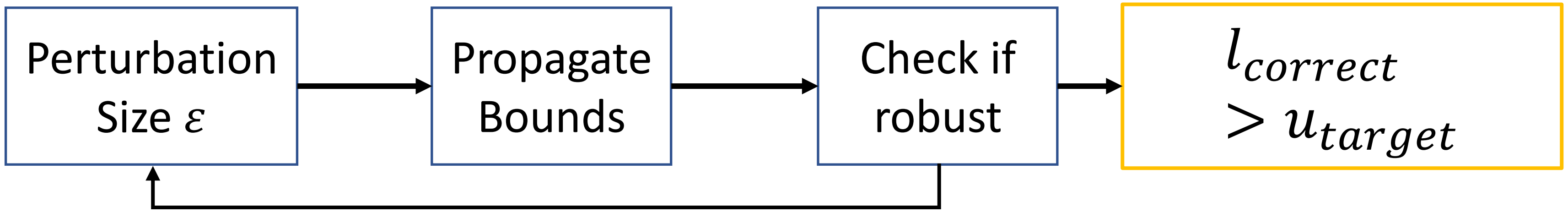
Efficient certified bound with activation bounds



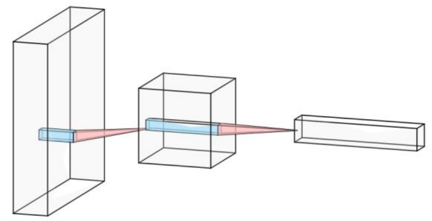
Efficient certified bound with activation bounds



Input



Efficient certified bound with activation bounds

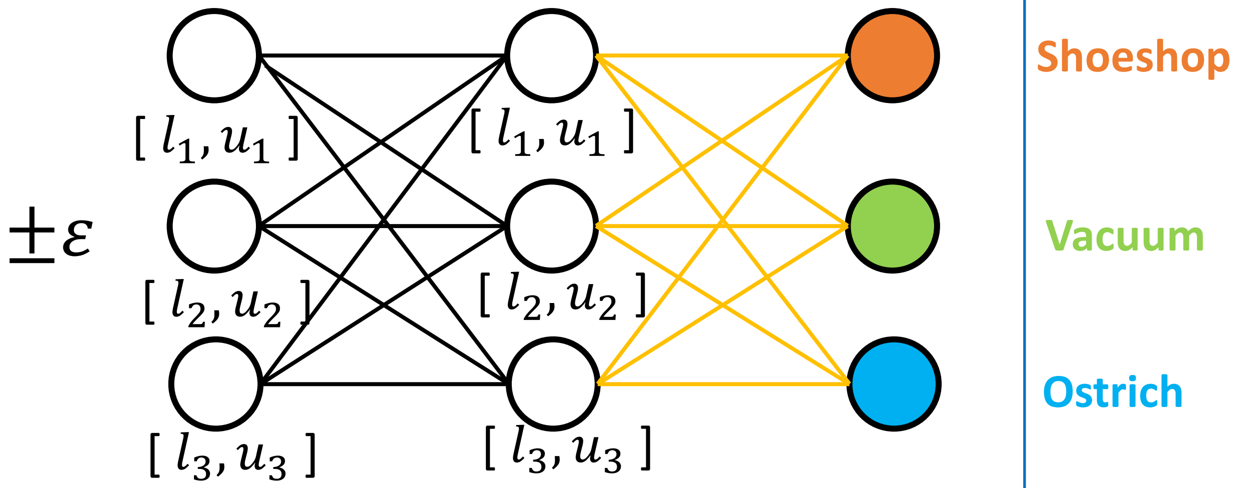


Trained CNN

Image x_0

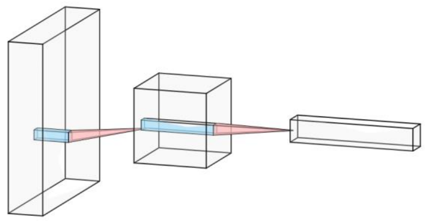
Input

$\|x - x_0\| \leq \varepsilon$



Perturbation Size ε → Propagate Bounds → Check if robust → $l_{correct} > u_{target}$

Efficient certified bound with activation bounds

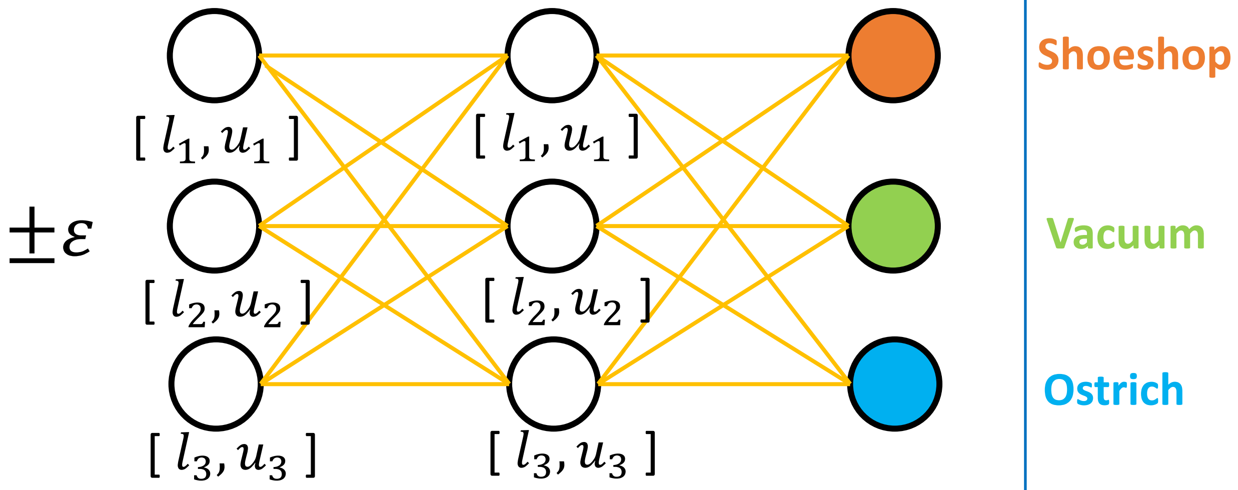


Trained CNN

Image x_0

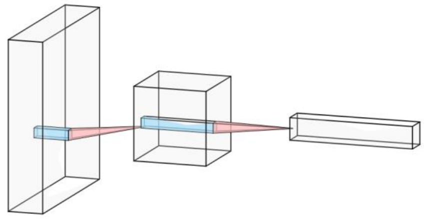
Input

$\|x - x_0\| \le \varepsilon$



Perturbation Size ε → Propagate Bounds → Check if robust → $l_{correct} > u_{target}$

Efficient certified bound with activation bounds

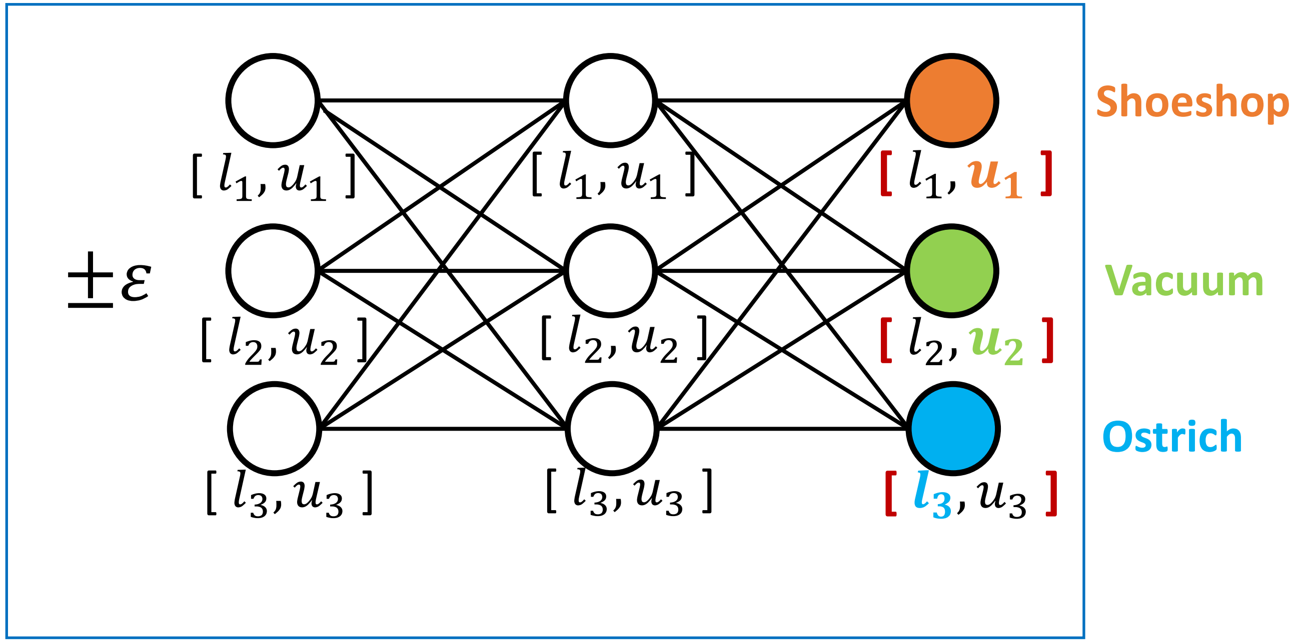


Trained CNN

Image x_0

Input

$\|x - x_0\| \leq \varepsilon$

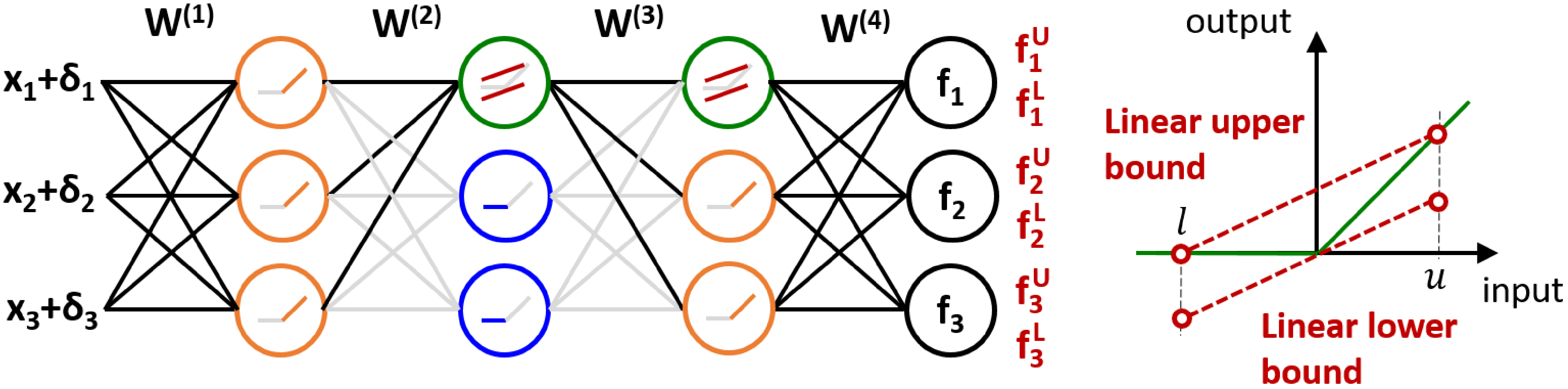


Perturbation Size ε → Propagate Bounds → Check if robust → $l_{correct} > u_{target}$

$l_{ostrich} > u_{vacuum}$

CROWN: certification with general activation functions

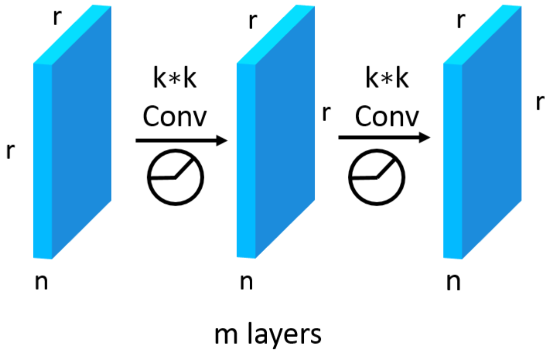
- How do we efficiently find the activation bounds for certification?



- By applying **adaptive linear** upper/lower bounds on the activation functions, we can derive explicit expression of m-layer neural network output given the input is constrained in an L_p-ball with radius ϵ. Thus a bisect ϵ can obtain max certified lower bound.

CNN-Cert represents bounds as convolutions

$f(x)$



$\|x - x_0\| \leq \varepsilon$

x_0= Original image
x= Perturbed image

$$L \leq f(x) \leq U$$

Fast-Lin[1]

$$L = Ax + B_L$$
$$U = Ax + B_U$$

CROWN[2]

$$L = A_L x + B_L$$
$$U = A_U x + B_U$$

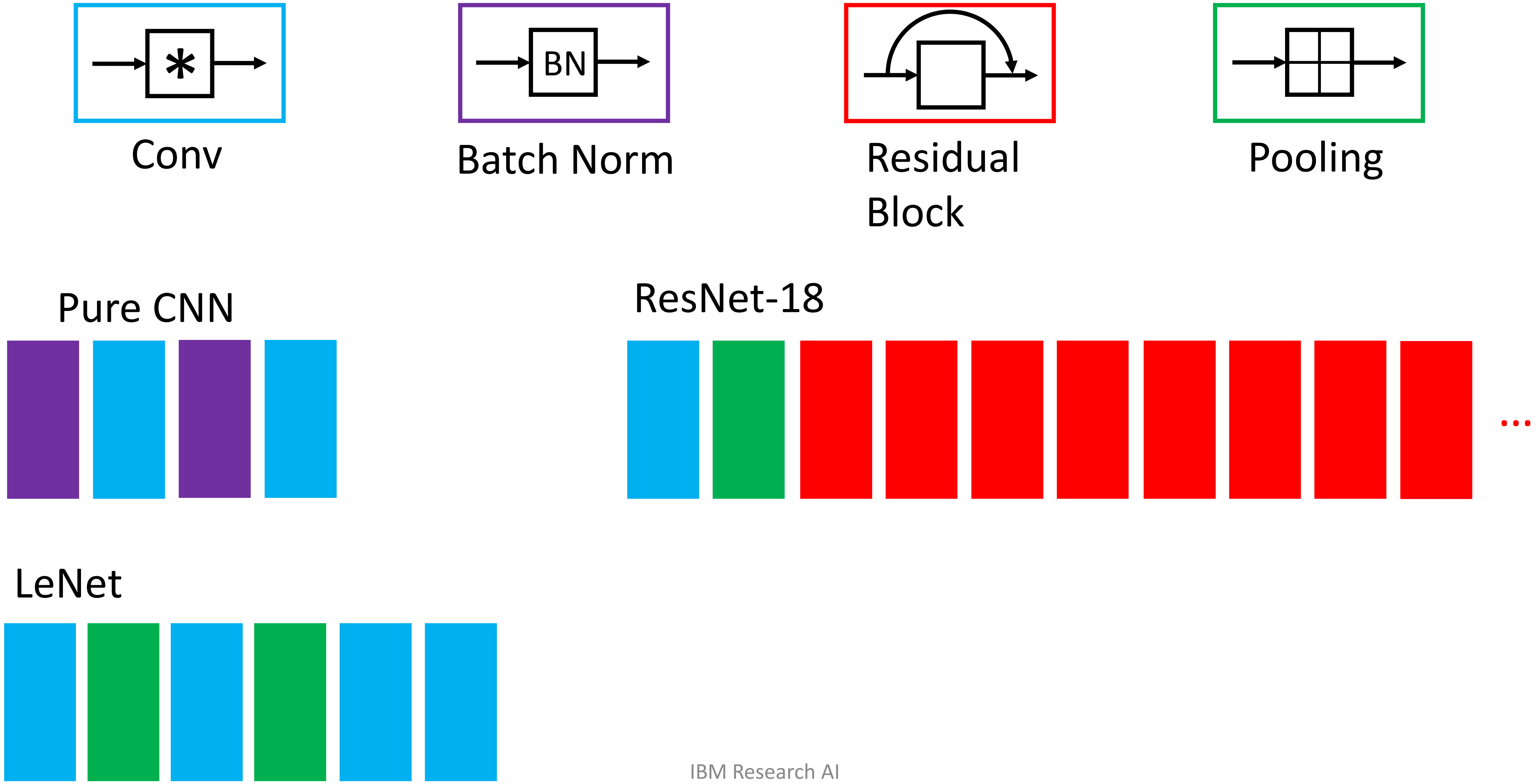
CNN-Cert

$$L = A_L * x + B_L$$
$$U = A_U * x + B_U$$

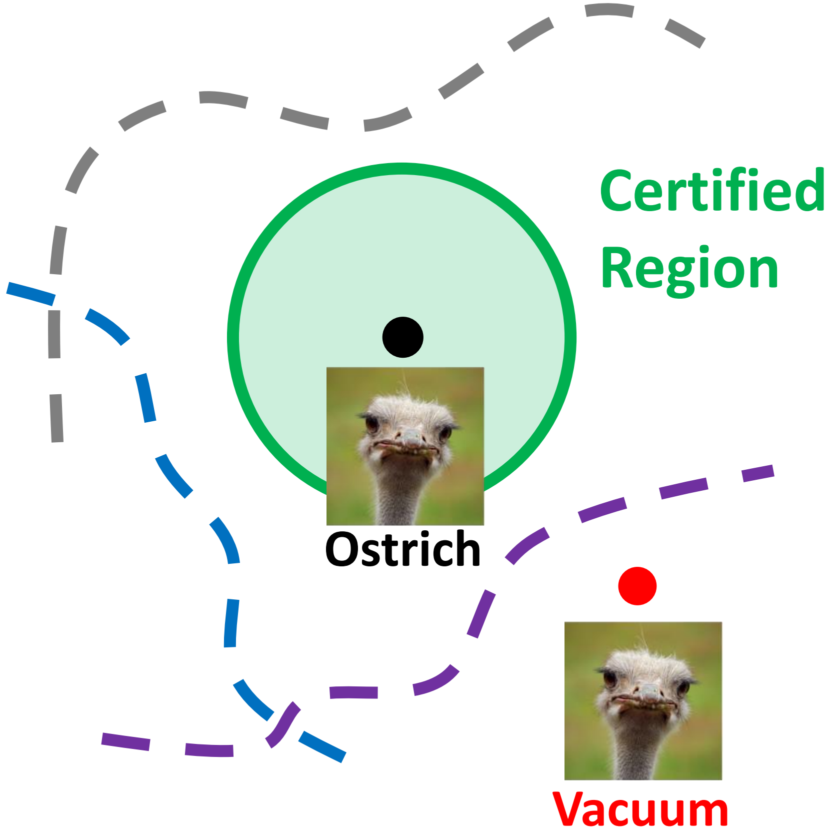
$*$ is the convolution operator

Towards Fast Computation of Certified Robustness for ReLU Networks, Tsui-Wei Weng*, Huan Zhang*, Hongge Chen, Zhao Song, Cho-Jui Hsieh, Duane Boning, Inderjit S. Dhillon and Luca Daniel, ICML 2018
Efficient Neural Network Robustness Certification with General Activation Functions, Huan Zhang*, Tsui-Wei Weng*, Pin-Yu Chen, Cho-Jui Hsieh and Luca Daniel, NeurIPS 2018
CNN-Cert: An Efficient Framework for Certifying Robustness of Convolutional Neural Networks, Akhilan Boopathy, Tsui-Wei Weng, Pin-Yu Chen, Sijia Liu, and Luca Daniel, AAAI 2019

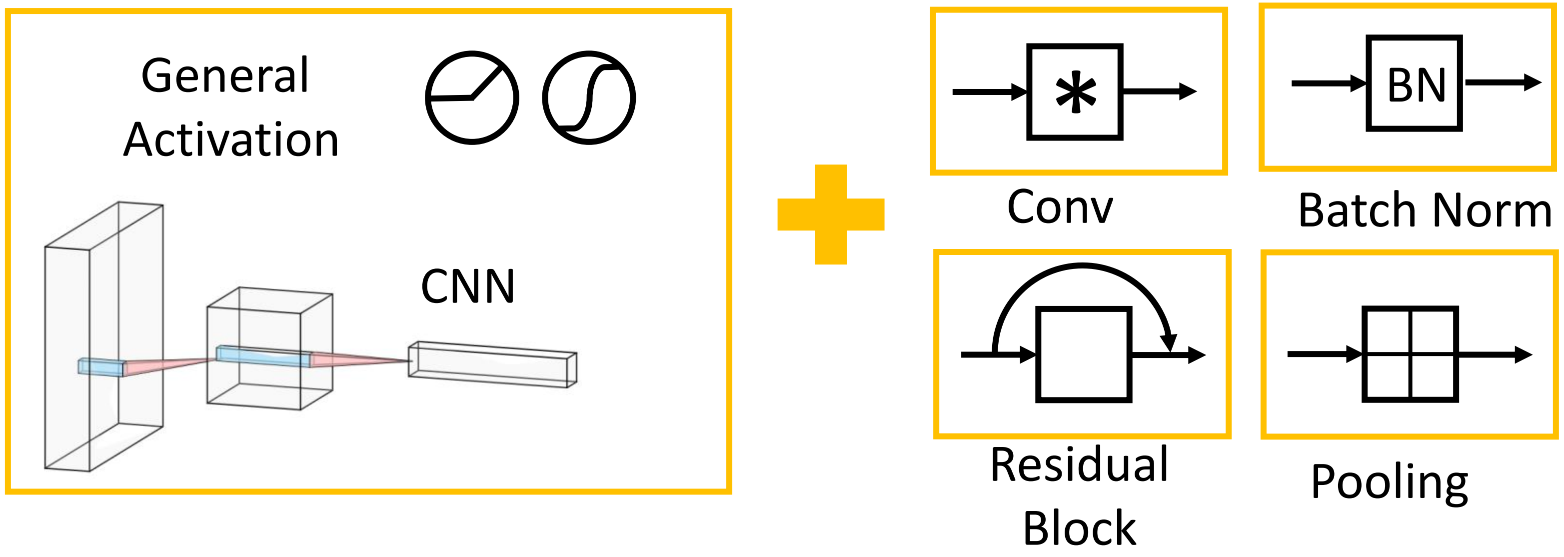
CNN-Cert supports various building blocks



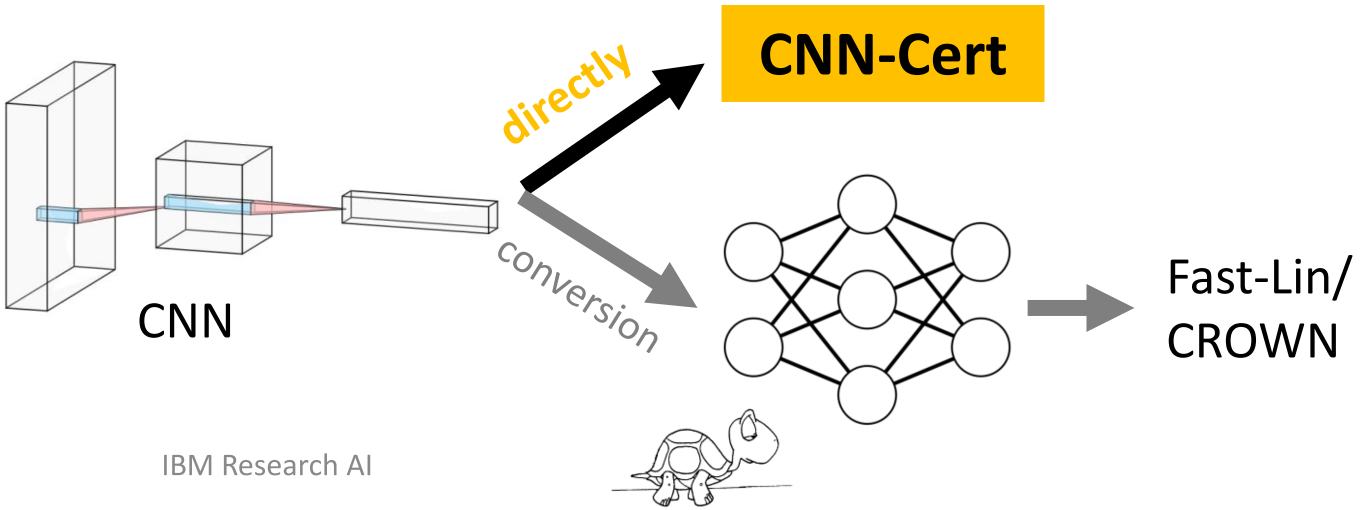
CNN-Cert finds a certified region of robustness



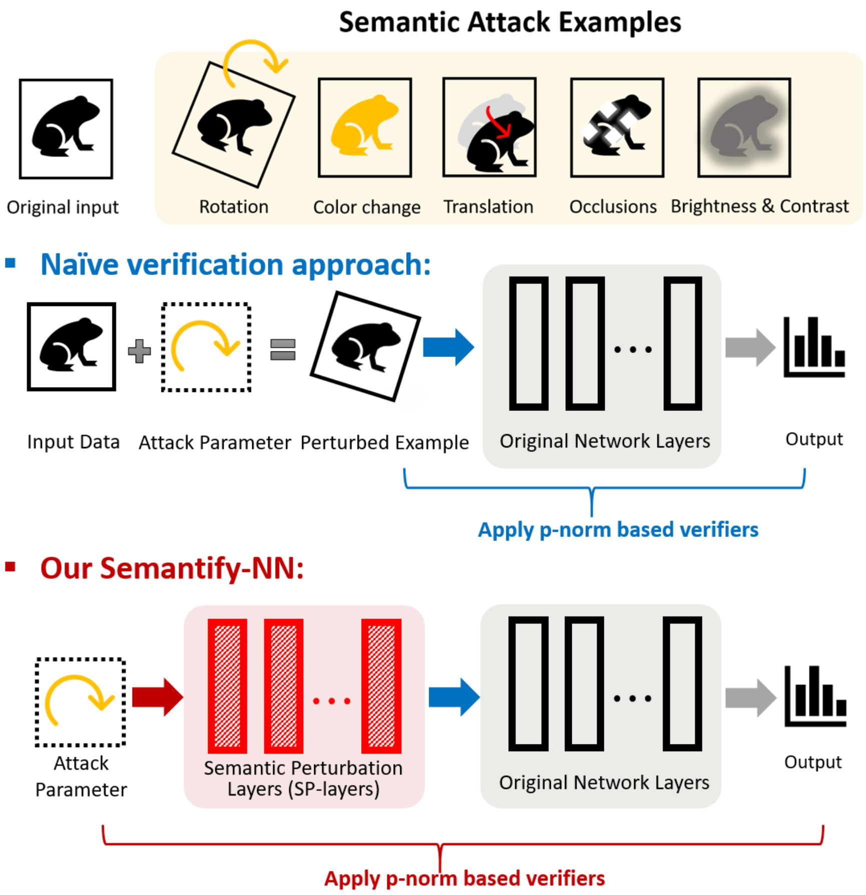
CNN-Cert is **general**...



...and **efficient**



Robustness Verification against Semantic Attacks



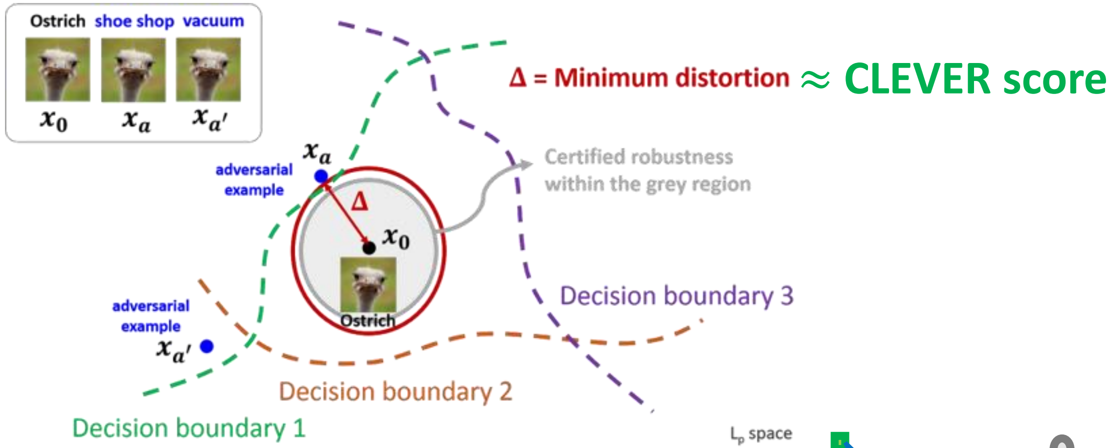
- Certificate of image rotation degree against prediction changes

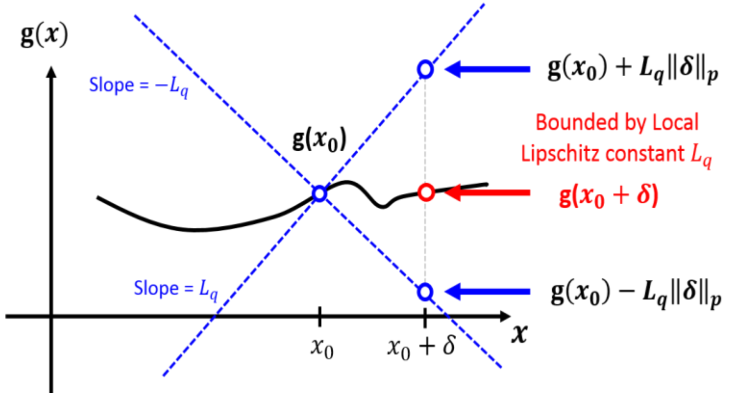
Network	Certified Bounds (degrees)				Attack (degrees)
	Number of Implicit Splits			**SPL + Refine**	Grid Attack
	1 implicit No explicit	5 implicit No explicit	10 implicit No explicit	100 implicit + explicit intervals of 0.5°	
Experiment (II): Rotations					
MNIST, MLP 2× 1024	0.627	1.505	2.515	46.24	51.42
MNIST, MLP 2× 1024 l_∞ adv	1.376	2.253	2.866	45.49	46.02
MNIST, CNN LeNet	0.171	0.397	0.652	43.33	48.00
CIFAR, MLP 5 × 2048	0.006	0.016	0.033	14.81	37.53
CIFAR, CNN 5 × 10	0.008	0.021	0.042	10.65	30.81
GTSRB, MLP 4 × 256	0.041	0.104	0.206	31.53	33.43

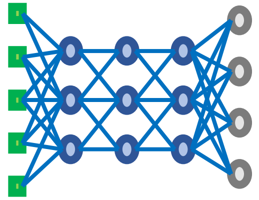
Jeet Mohapatra, Tsui-Wei (Lily) Weng, Pin-Yu Chen, Sijia Liu, and Luca Daniel, "Towards Verifying Robustness of Neural Networks Against Semantic Perturbations," *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020

CLEVER: a tale of two approaches

- An attack-independent, model-agnostic robustness metric that is efficient to compute
- Derived from theoretical robustness analysis for verification of neural networks: Cross Lipschitz Extreme Value for nEtwork Robustness
- Use of extreme value theory for efficient estimation of minimum distortion
- Scalable to large neural networks
- Open-source codes: https://github.com/IBM/CLEVER-Robustness-Score







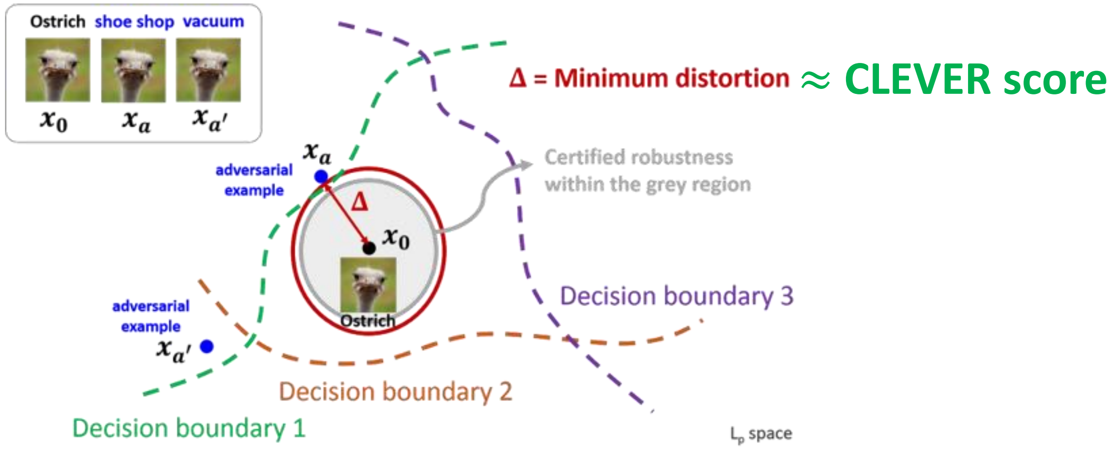
input-output perturbation analysis of neural net

Evaluating the Robustness of Neural Networks: An Extreme Value Theory Approach, Tsui-Wei Weng*, Huan Zhang*, Pin-Yu Chen, Jinfeng Yi, Dong Su, Yupeng Guo, Cho-Jui Hsieh, and Luca Daniel, ICLR 2018
On Extensions of CLEVER: a Neural Network Robustness Evaluation Algorithm, Tsui-Wei Weng*, Huan Zhang*, Pin-Yu Chen, Aurelie Lozano, Cho-Jui Hsieh, and Luca Daniel, GlobalSIP 2018

CLEVER way for Adversarial Robustness Evaluation

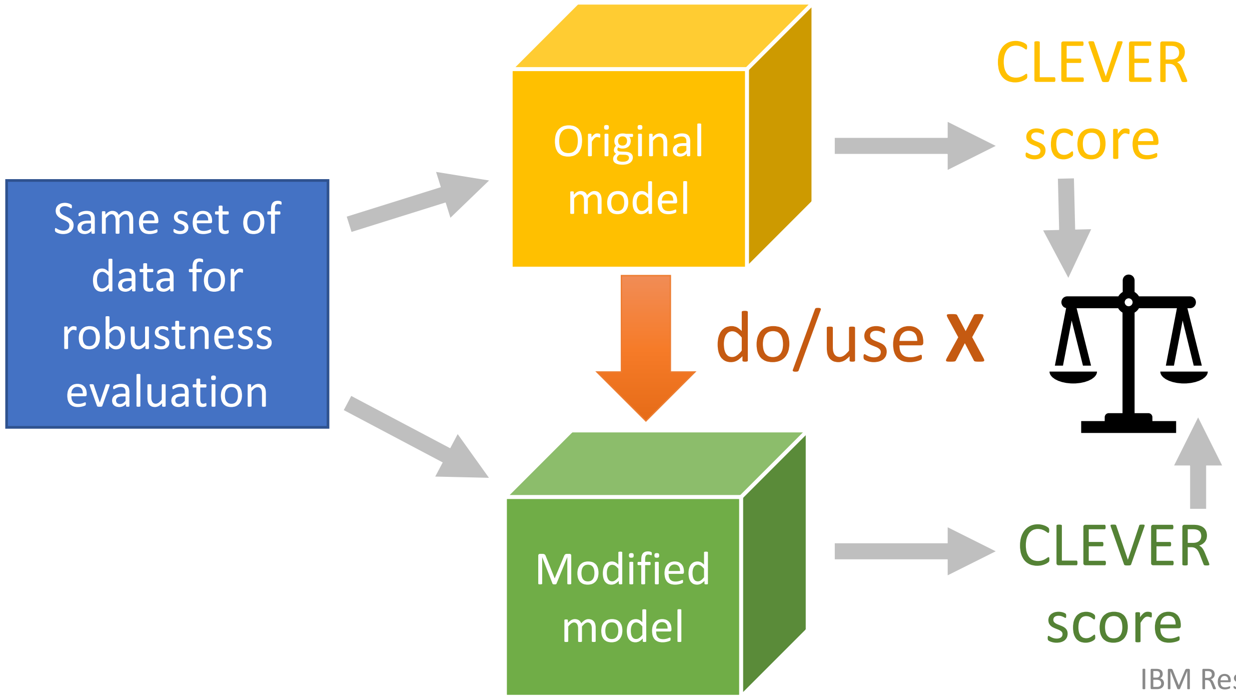
An attack-independent, model-agnostic robustness metric that is efficient to compute

Evaluating the Robustness of Neural Networks: An Extreme Value Theory Approach, Tsui-Wei Weng*, Huan Zhang*, Pin-Yu Chen, Jinfeng Yi, Dong Su, Yupeng Guo, Cho-Jui Hsieh, and Luca Daniel, ICLR 2018



Before-After robustness comparison

- Will my model become more robust if I do/use **X**?



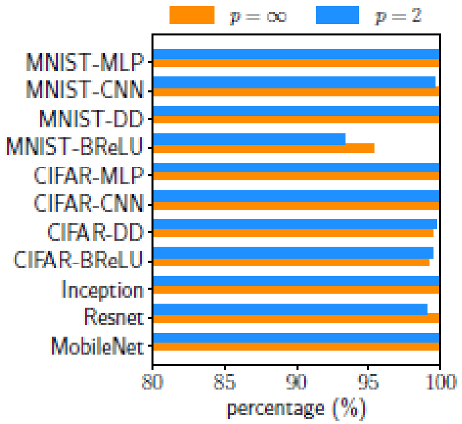
Other use cases

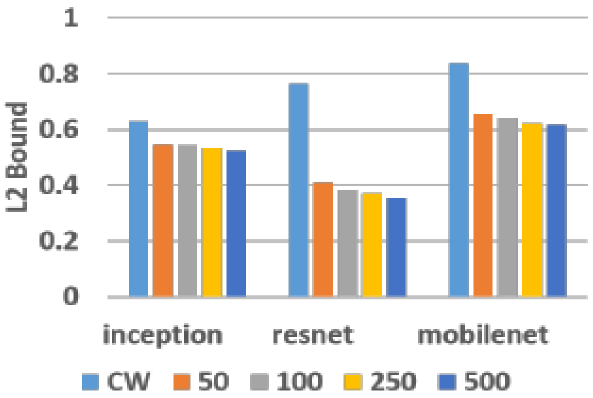
- Characterize the behaviors and properties of adversarial examples
- Hyperparameter selection for adversarial attacks and defenses
- Reward-driven model robustness improvement

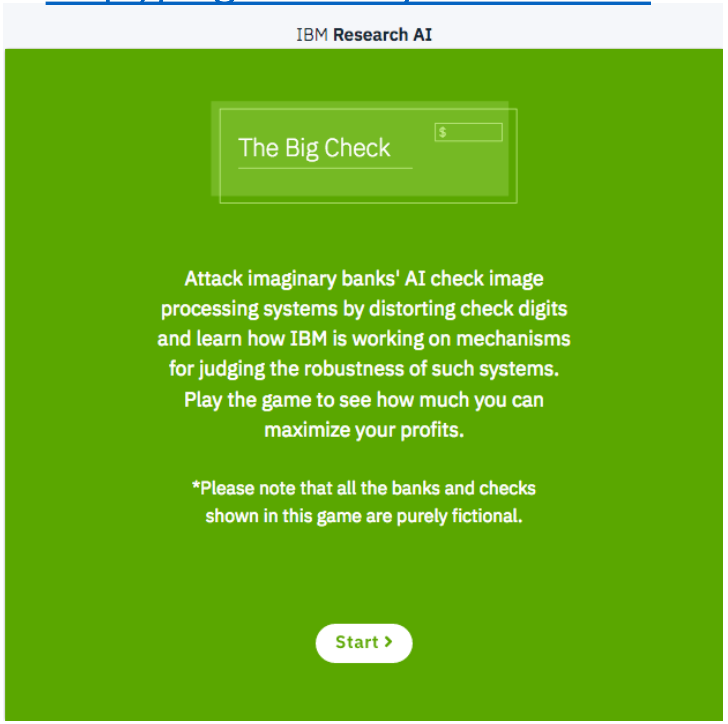
http://bigcheck.mybluemix.net

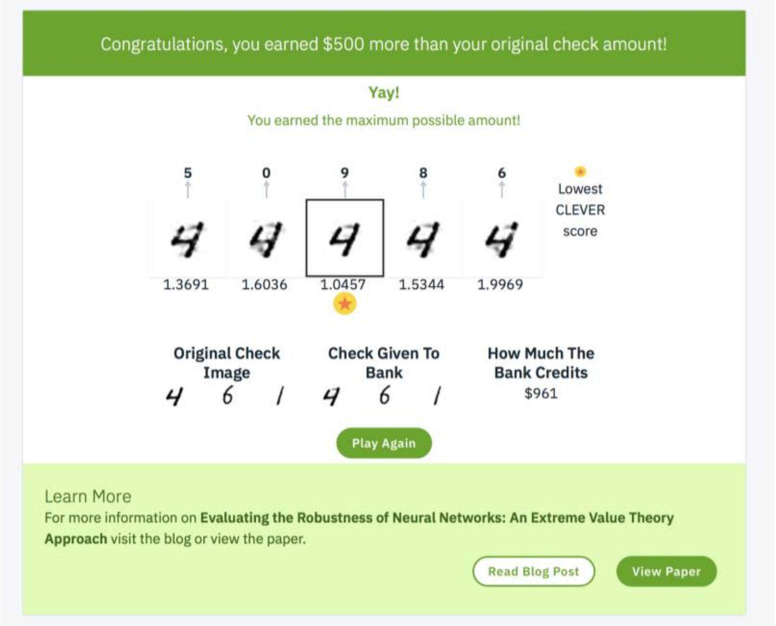
Examples of CLEVER

- CLEVER enables robustness comparison between <u>different</u>
- ❑ Threat models
- ❑ Datasets
- ❑ Neural network architectures
- ❑ Defense mechanisms



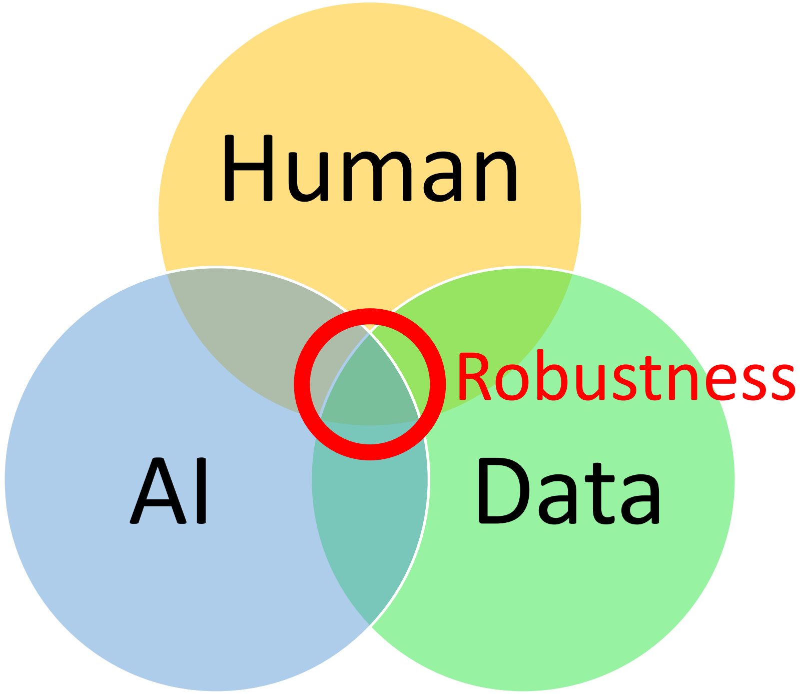




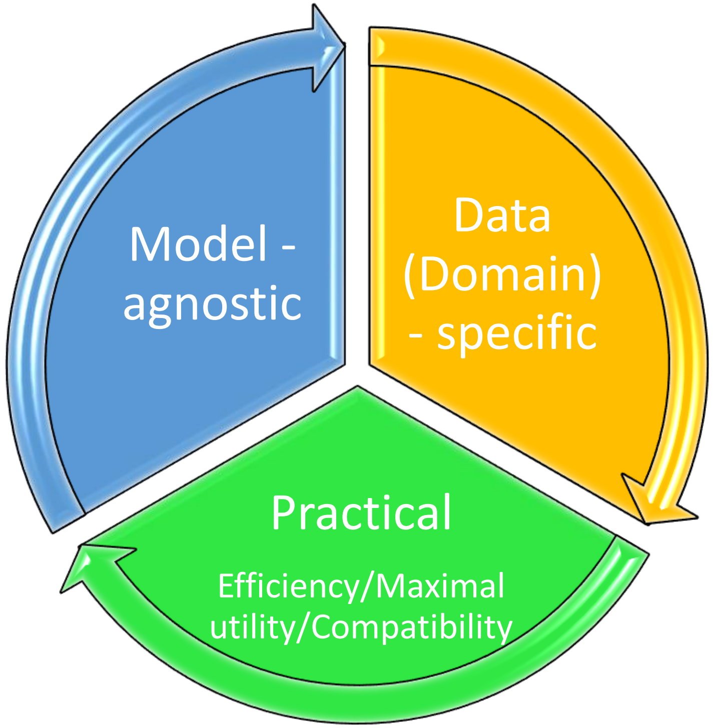


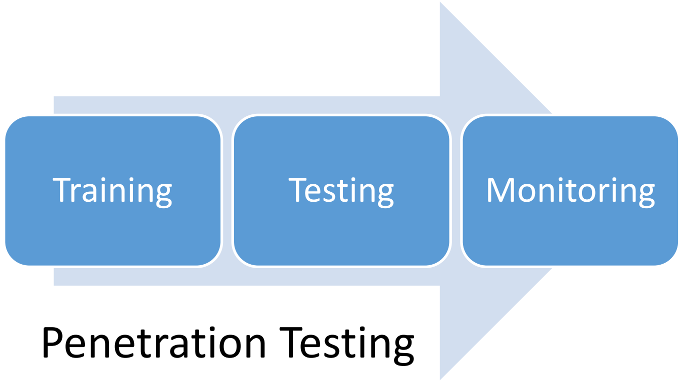
Take-aways

- Adversarial robustness is a new AI standard toward trustworthy ML
  - ❑Robustness does not come for free: adversarial examples exist in digital space, physical world, and different domains
  - ❑High accuracy $\neq$ Good robustness
  - ❑Arms race: adversary-aware AI v.s. AI for adversary
- How to evaluate and improve model robustness?
  - ❑Various attack threat models and taxonomy
  - ❑Incorporate domain knowledge, attack-agnostic defense
  - ❑Scalable and efficient robust training & verification
- Adversarial machine learning beyond attacks and defenses
  - ❑Model reprogramming
- Join us for the exciting journey!
- Twitter: @pinyuchenTW



Roadmap toward Holistic Adversarial Robustness





Attack (Bug Finding)

- In-house **sensitivity and reliability tests** for developed models
- Generate prediction-evasive examples (per user constraints)
- Customize to model deployment conditions (e.g. cloud APIs)

Defense (Model Hardening)

- **Detecting and mitigating** potential adversarial threats
- **Plug-and-play** model patching for a given model
- Landscape exploration: model fix and cleaning

Verification (Model Certificate)

- This model is certified to be **attack-proof** up to a certain level
- Quantifiable metric for certified robustness
- AI standards, governance, and law regulation

Applications to AI (Model Boosting)

- Data augmentation
- **Model reprogramming**: data-efficient transfer learning
- Model watermarking

Online Resources for Adversarial Robustness

- J. Z. Kolter and A. Madry: Adversarial Robustness - Theory and Practice (NeurIPS 2018 Tutorial)
- Pin-Yu Chen: Adversarial Robustness of Deep Learning Models (ECCV 2020 Tutorial)
- Pin-Yu Chen and Sijia Liu: Zeroth Order Optimization: Theory and Applications to Deep Learning (CVPR 2020 Tutorial)
- Pin-Yu Chen and Sayak Paul: Practical Adversarial Robustness in Deep Learning: Problems and Solutions (CVPR 2021 Tutorial)

cleverhans

Adversarial Robustness Toolbox (ART v0.10.0)

Foolbox

Sample Surveys for Adversarial Robustness

Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning

Battista Biggio[a,b,*], Fabio Roli[a,b]

[a] *Department of Electrical and Electronic Engineering, University of Cagliari, Italy*
[b] *Pluribus One, Cagliari, Italy*

On Evaluating Adversarial Robustness

Nicholas Carlini[1], Anish Athalye[2], Nicolas Papernot[1], Wieland Brendel[3], Jonas Rauber[3], Dimitris Tsipras[2], Ian Goodfellow[1], Aleksander Mądry[2], Alexey Kurakin[1 *]

[1] Google Brain [2] MIT [3] University of Tübingen

The Robustness of Deep Networks

A geometrical perspective

Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard

On Adaptive Attacks to Adversarial Example Defenses

Florian Tramèr* Stanford University
Nicholas Carlini* Google Brain
Wieland Brendel* University of Tübingen
Aleksander Mądry MIT

Adversarial Learning Targeting Deep Neural Network Classification: A Comprehensive Review of Defenses Against Attacks

Publisher: IEEE Cite This PDF

3 Author(s) David J. Miller ; Zhen Xiang ; George Kesidis **View All Authors**

- Book on "Adversarial Machine Learning" authored by Cho-Jui Hsieh@UCLA and Pin-Yu Chen, to appear in 2022

Making AI model Robust is truly ART

Adversarial Robustness Toolbox (ART)

External: https://github.com/IBM/adversarial-robustness-toolbox

- Python library, 7K lines of code
- State-of-the-art attacks, defences and robustness metrics

Load ART modules →

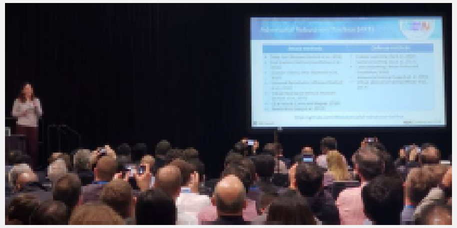
Load classifier model (Keras, TF, PyTorch etc) →

Perform attack →

Evaluate robustness →

```
from keras.datasets import mnist
from keras.models import load_model

from art.attacks import CarliniL2Attack
from art.classifier import KerasClassifier
from art.metrics import loss_sensitivity

# Load data
(_, _), (x_test, y_test) = mnist.load_data()

# Load model and build classifier
model = load_model('my_favorite_keras_model.h5')
classifier = KerasClassifier((0, 1), model)

# Perform attack
attack = CarliniL2Attack(classifier)
adv_x_test = attack.generate(x_test)

# Compute metrics on model robustness
print(loss_sensitivity(classifier, x_test))
```

Open-source release @ RSA 2018:



~ 3.5K+ views of IBM blogs
~ 100+ news outlets covering release
~ 1.3M+ Social Media potential impressions
~ 5K+ views of GitHub repo

siliconANGLE
Attackers can fool AI programs. Here's how developers can fight back

ZDNet
IBM launches open-source library for securing AI systems
The framework-agnostic software library contains attacks, defenses, and benchmarks for securing artificial intelligence systems.

IBM ENTWICKELT WERKZEUGE GEGEN HACKERANGRIFFE DURCH "BÖSE" KI

IBM、AIシステムを保護するオープンソースライブラリ「Adversarial Robustness Toolbox」

Выпущена Adversarial Robustness Toolbox, открытая библиотека от IBM для защиты ИИ

Adversarial Robustness Toolbox : IBM propose une boite à outils open source pour sécuriser l'intelligence artificielle

23-04-2018 | door: Witold Kepinski
IBM Adversarial Robustness Toolbox beschermt tegen kwaadaardige AI

Evasion attacks	Evasion defenses	Poisoning detection	Robustness metrics
• FGSM • JSMA	• Feature squeezing • Spatial smoothing	• Detection based on clustering activations	• CLEVER • Empirical robustness

3rd Workshop on Adversarial Learning Methods for Machine Learning and Data Mining @ KDD 2021 (virtual workshop)

Call for Papers

Organizers & Committee

- **One Best Paper Awards and Two Rising Star Awards are sponsored by MIT-IBM Watson AI Lab with cash prizes ($500 each)!**
- Co-located conference: KDD 2021 (virtual conference)
- Workshop Date and time: TBA
- Organizers: Pin-Yu Chen (IBM Research), Cho-Jui Hsieh (UCLA), Bo Li (UIUC), SIjia Liu (Michigan State University)
- Paper submission Deadline: May 20th, 2021
- Notification Date: June 10th, 2021
- Submission Site: CMT
- Paper submission format: ACM template, **4 pages** excluding references and supporting materials. The authors can choose to anonymize the author information during submission (but not required to do so).

IBM | AI Research   Research areas   Publications   Experiments   Careers   Blog

Trusted AI

IBM Research is building and enabling AI solutions people can trust

As AI advances, and humans and AI systems increasingly work together, it is essential that we trust the output of these systems to inform our decisions. Alongside policy considerations and business efforts, science has a central role to play: developing and applying tools to wire AI systems for trust. IBM Research's comprehensive strategy addresses multiple dimensions of trust to enable AI solutions that inspire confidence.

Robustness

We are working to ensure the security and reliability of AI systems by exposing and fixing their vulnerabilities: identifying new attacks and defense, designing new adversarial training methods to strengthen against attack, and developing new metric to evaluate robustness.

View publications

Fairness

To encourage the adoption of AI, we must ensure it does not take on and amplify our biases. We are creating methodologies to detect and mitigate bias through the life cycle of AI applications.

View publications

Explainability

Knowing how an AI system arrives at an outcome is key to trust, particularly for enterprise AI. To improve transparency, we are researching local and global interpretability of models and their output, training for interpretable models and visualization of information flow within models, and teaching explanations.

View publications

Lineage

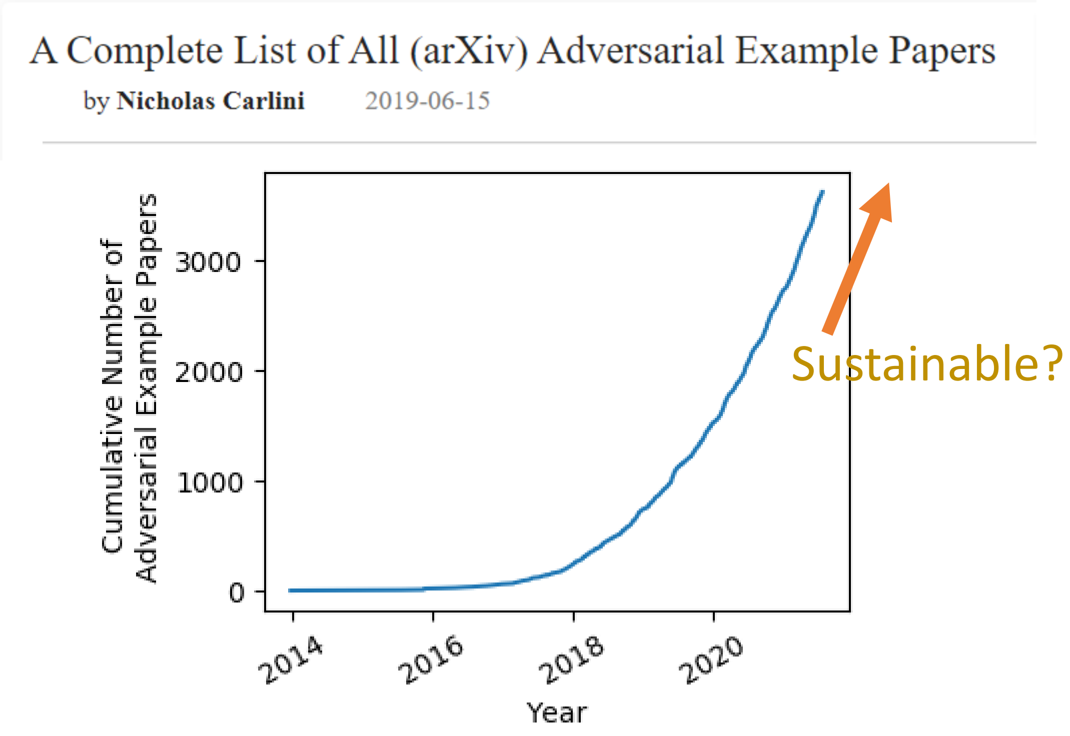
Lineage services can infuse trust in AI systems by ensuring all their components and events are trackable. We are developing services like instrumentation and event generation, scalable event ingestion and management, and efficient lineage query services to manage the complete lifecycle of AI systems.

View publications

IBM Research AI

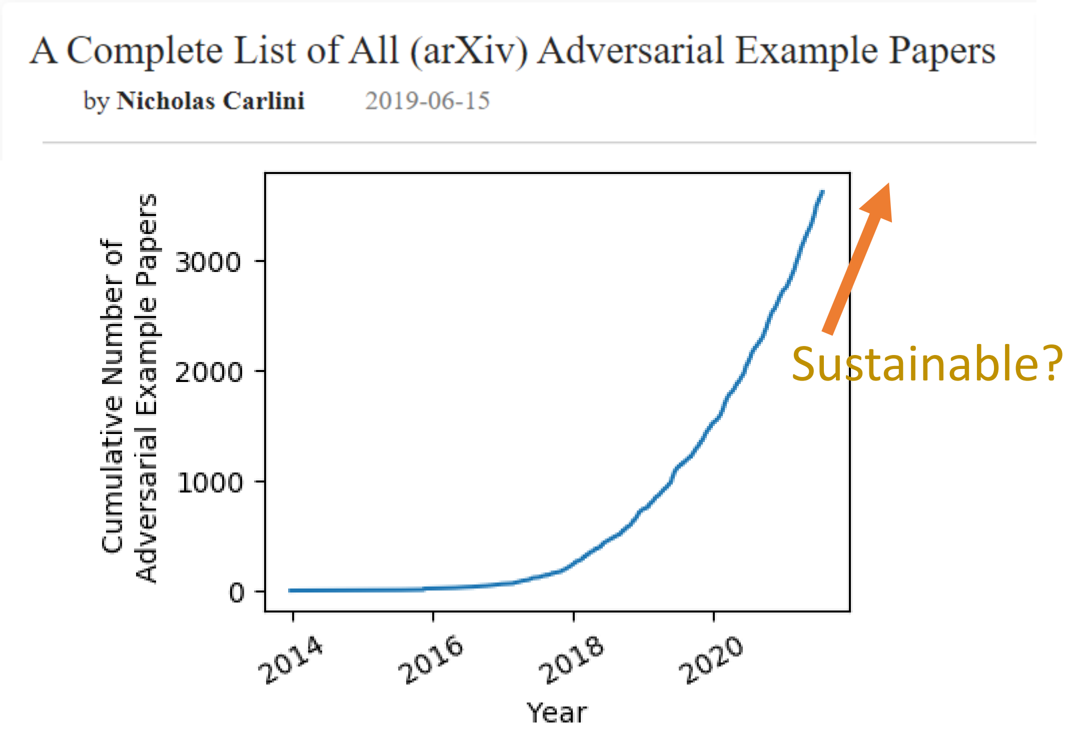
Trends I observed in Adversarial Machine Learning

- Attack:
  - Adversarial attack on [Task]
  - Black-box adversarial attack on [Task]
  - Hard-label black-box adversarial attack on [Task]
  - Efficient adversarial attack for [Perturbation Norm]
- Defense:
  - Defending against adversarial attacks using [Method]
  - Detecting adversarial examples using [Method]
  - Certified robustness for [Task]/[Norm]
  - Adversarial training using [Technique]
- Reflection:
  - All empirical defenses are vulnerable
  - How practical is the threat model? (e.g. unrestricted adversarial examples)
  - Intriguing properties of [New Network Architecture]
  - Tradeoff between adversarial robustness and [Factor] (e.g. privacy, fairness, interpretability)
  - Hardness of adversarial ML: optimization and generalization



Trends I observed in Adversarial Machine Learning

- Attack:
  - Adversarial attack on [Task]
  - Black-box adversarial attack on [Task]
  - Hard-label black-box adversarial attack on [Task]
  - Efficient adversarial attack for [Perturbation Norm]
- Defense:
  - Defending against adversarial attacks using [Method]
  - Detecting adversarial examples using [Method]
  - Certified robustness for [Task]/[Norm]
  - Adversarial training using [Technique]
- Reflection:
  - All empirical defenses are vulnerable
  - How practical is the threat model? (e.g. unrestricted adversarial examples)
  - Intriguing properties of [New Network Architecture]
  - Tradeoff between adversarial robustness and [Factor] (e.g. privacy, fairness, interpretability)
  - Hardness of adversarial ML: optimization and generalization



Acknowledgement

- My incredible collaborators (IBM Research, MIT, UCLA, North Eastern Univ, UIUC, Georgia Tech, Univ Minnesota, RPI, and many others)
- MIT-IBM Watson AI Lab https://mitibmwatsonailab.mit.edu/
- RPI-IBM AI Research Collaboration https://airc.rpi.edu/
- IBM AI Horizon Network: https://www.research.ibm.com/artificial-intelligence/horizons-network/
- IBM Trusted AI Group: Payel Das, Saska Mojsilovic
- IBM AI-Security Group
- IBM Big Check Demo Group

❑ Personal Website: www.pinyuchen.com

❑ Twitter: pinyuchen.tw

Now is the time to query me for questions!

- Pin-Yu Chen
- pin-yu.chen@ibm.com
- www.pinyuchen.com
- Twitter: @pinyuchenTW

Q&A