How CLEVER is your neural network?

Robustness evaluation against adversarial examples

Pin-Yu Chen

IBM Research AI

O’Reilly AI Conference @ London 2018

Label it!

Label it! AI model says:

ostrich

How about this one?

Surprisingly, AI model says:

shoe shop

ostrich | safe | shoe shop | vacuum

What is wrong with this AI model?

- This model is one of the BEST image classifier using neural networks

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

Adversarial examples: the evil doublegangers

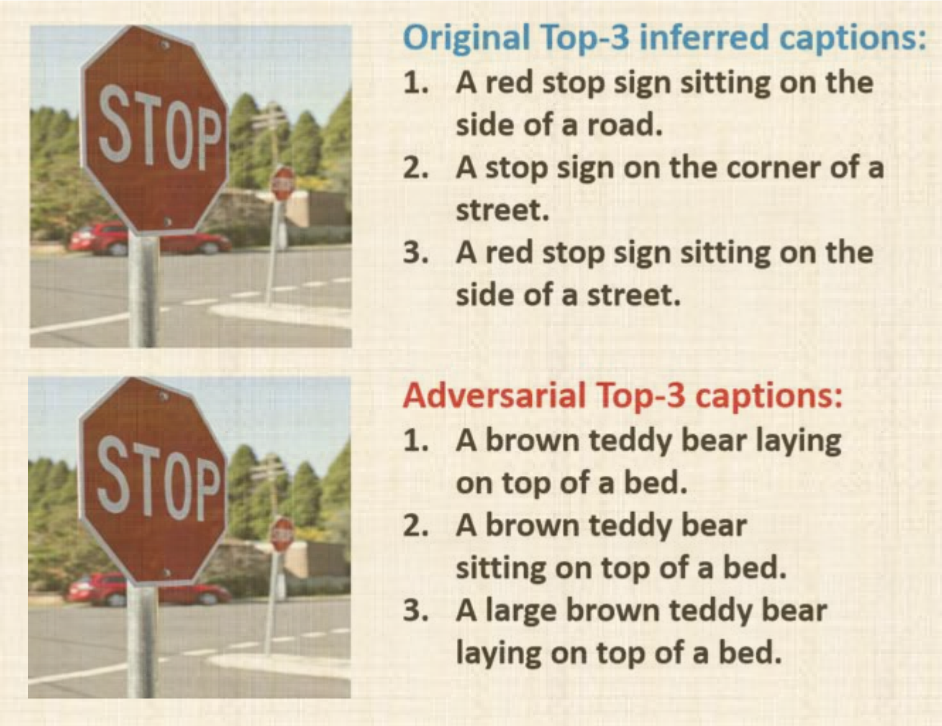
source: Google Images

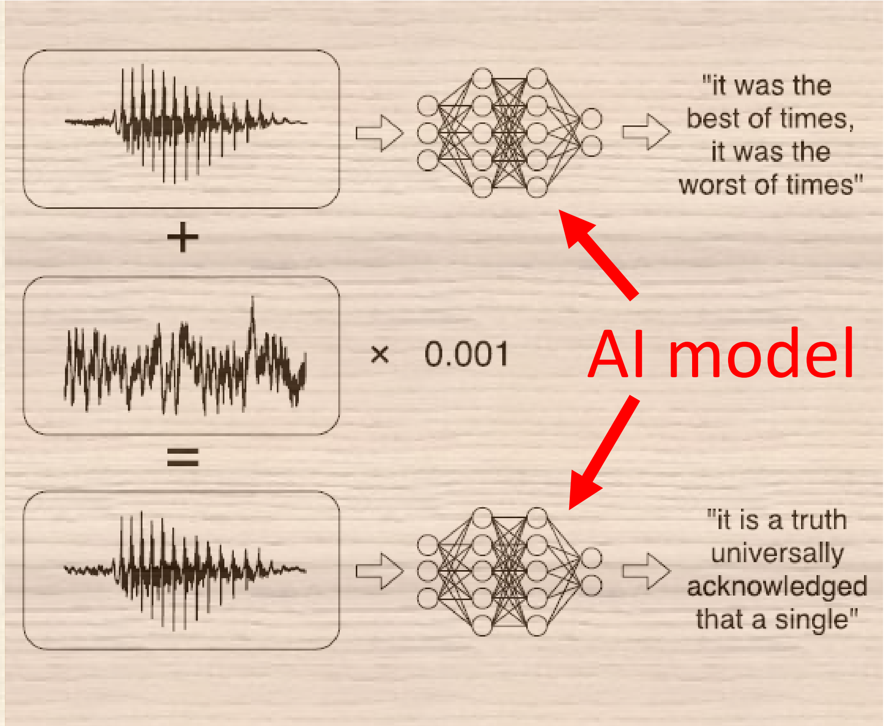
Why do adversarial examples matter?

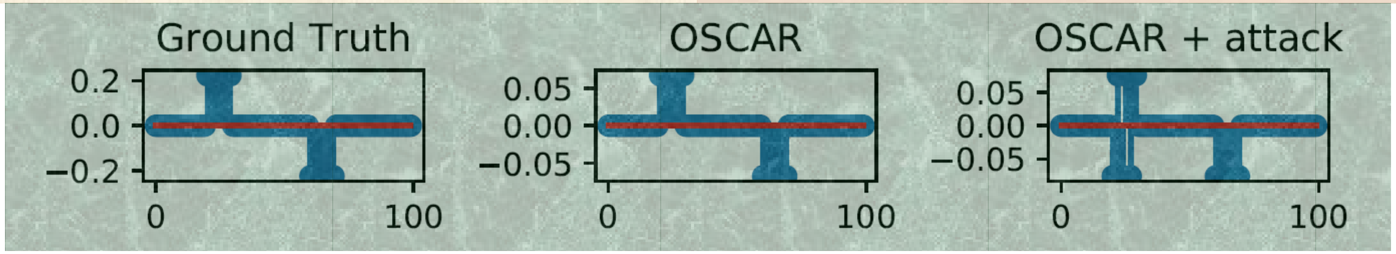
- Adversarial attacks on an AI model deployed at test time (aka evasion attacks)

Adversarial examples in different domains

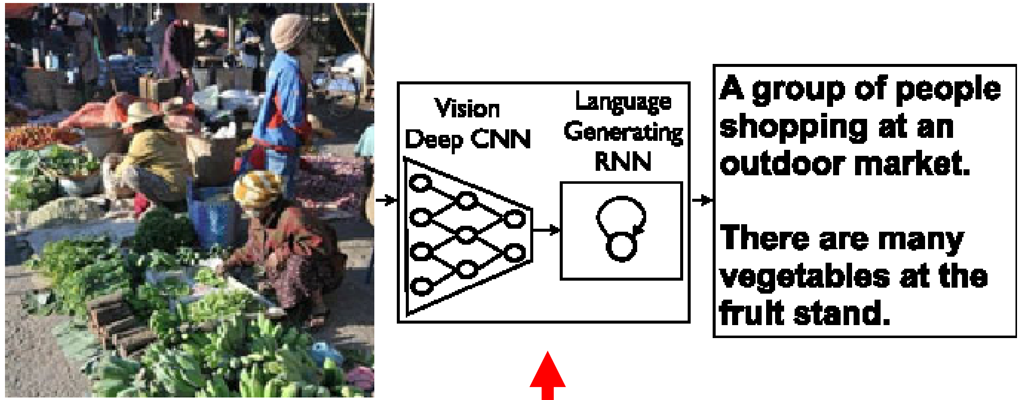
- Images
- Videos
- Texts
- Speech/Audio
- Data analysis
- Electronic health records
- Malware
- Online social network
- and many others







Adversarial examples in image captioning



AI model

Input: image

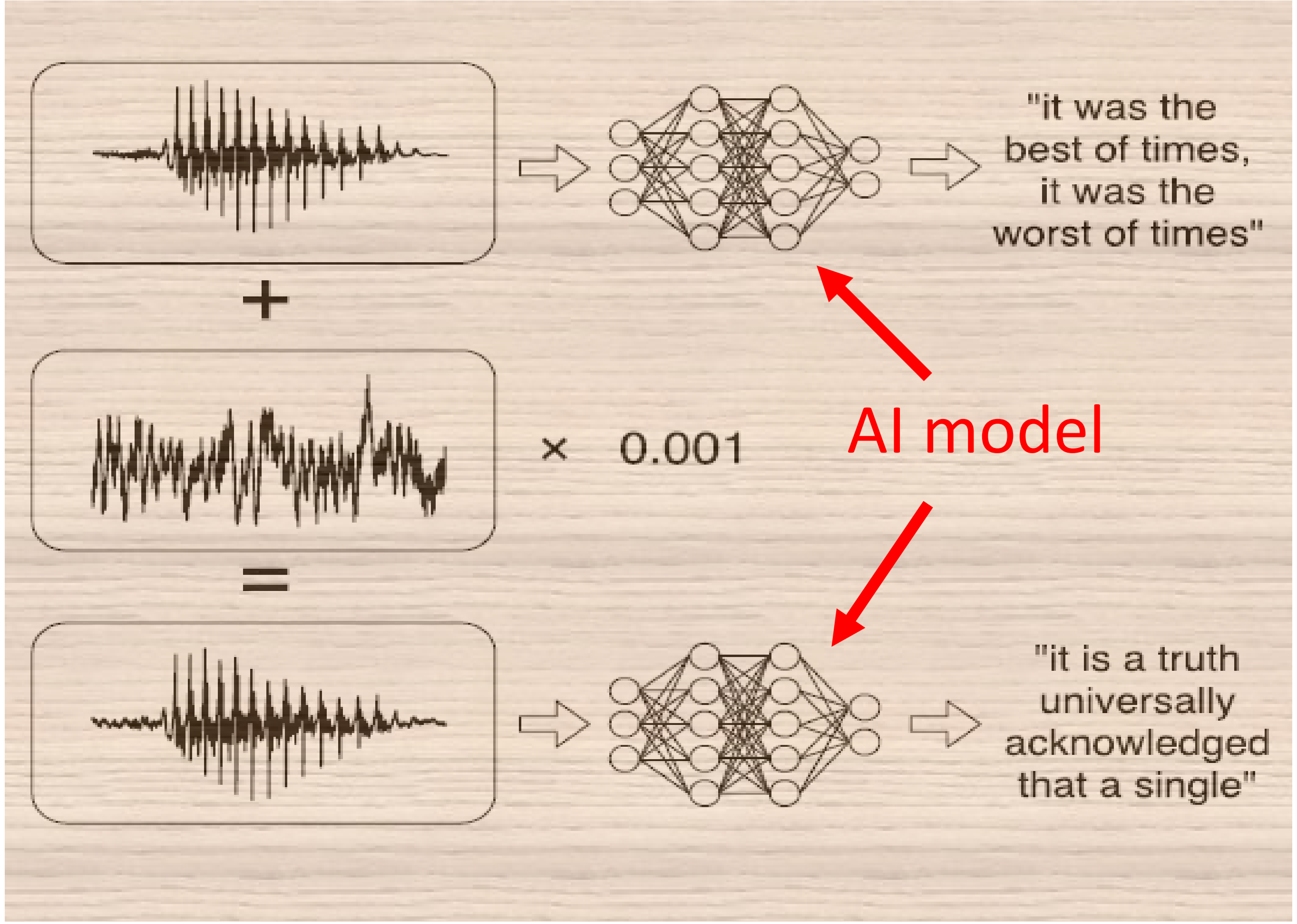
Output: caption

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

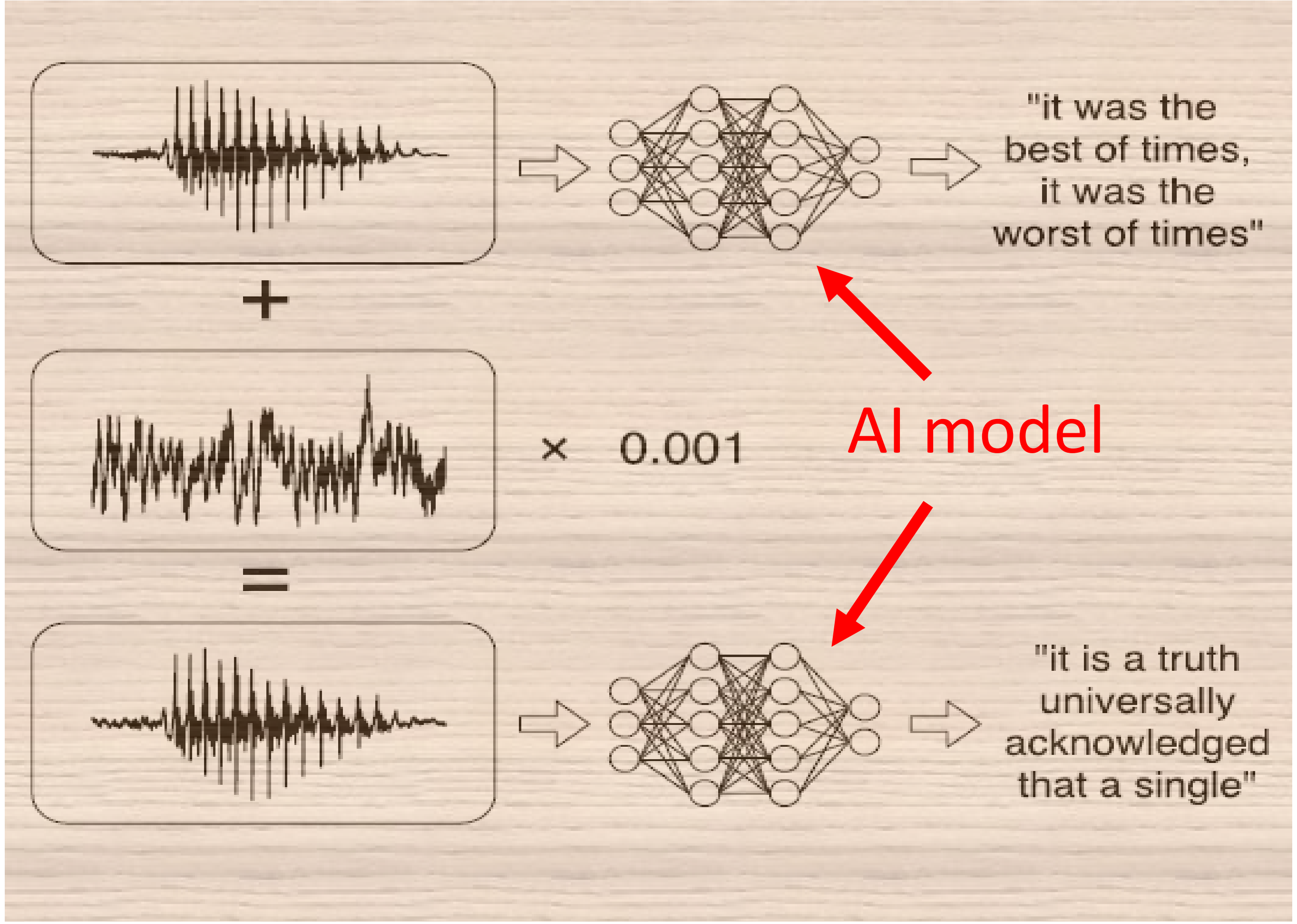


without the dataset the article is useless

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



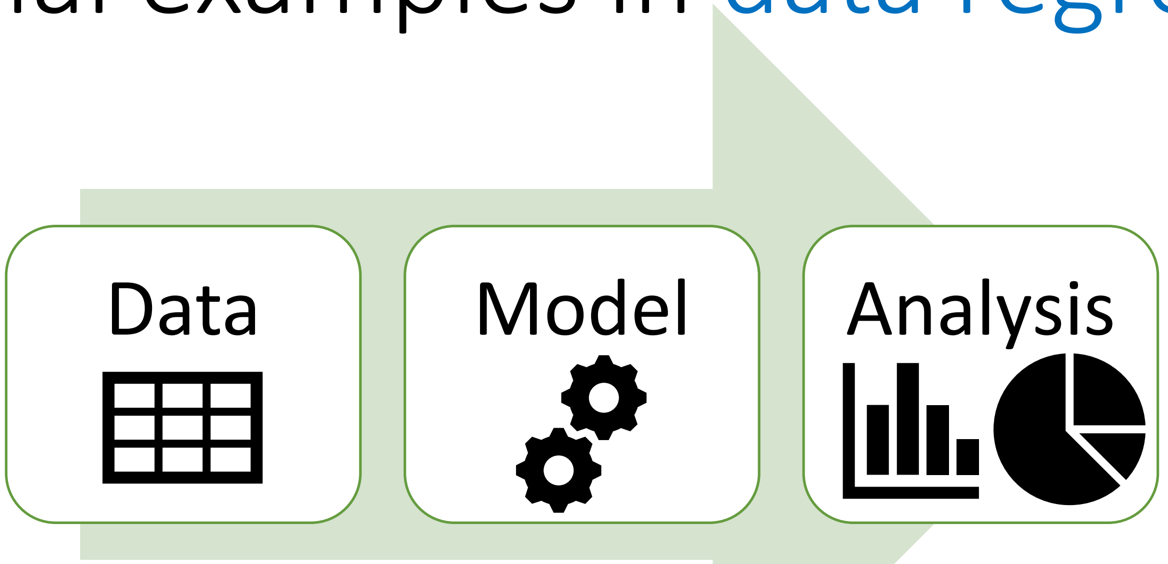
without the dataset the article is useless

What did your hear?

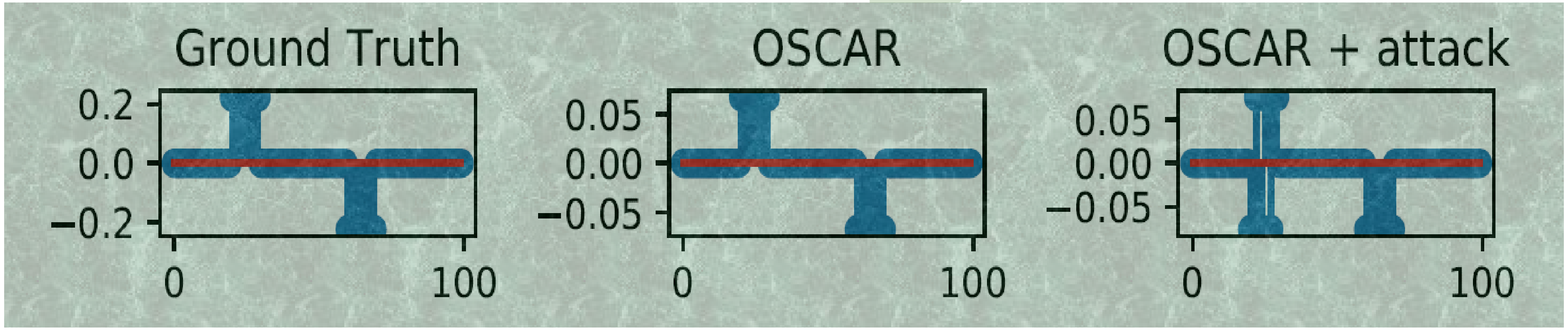
okay google browse to evil.com

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



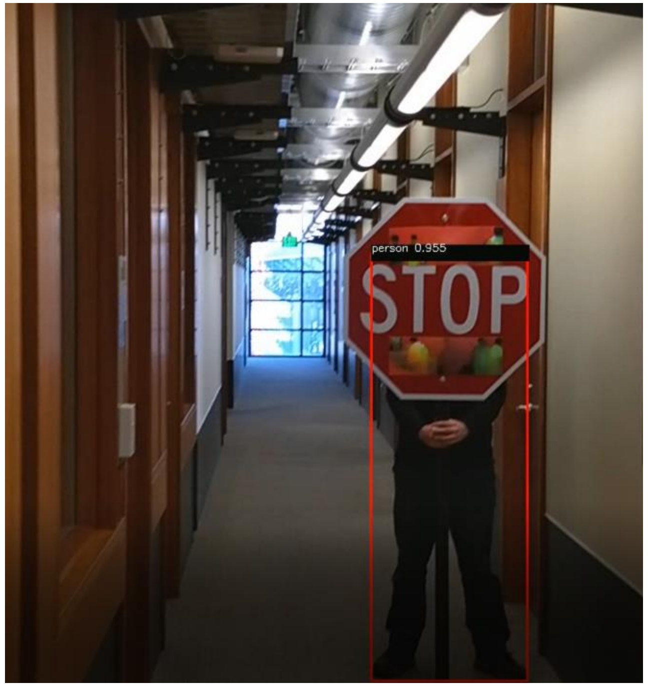
Factor identification



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 physical world

- Real-time traffic sign detector



- 3D-printed adversarial turtle

classified as turtle   classified as rifle   classified as other

Artificial Intelligence CONFERENCE

PUT AI TO WORK
APRIL 29-30, 2018: TRAINING
APRIL 30-MAY 2, 2018: TUTORIALS & CONFERENCE
NEW YORK, NY

SCHEDULE SPEAKERS EVENTS SPONSORS VENUE/HOTEL ABOUT RESOURCES ACCOUNT

Fooling neural networks in the physical world

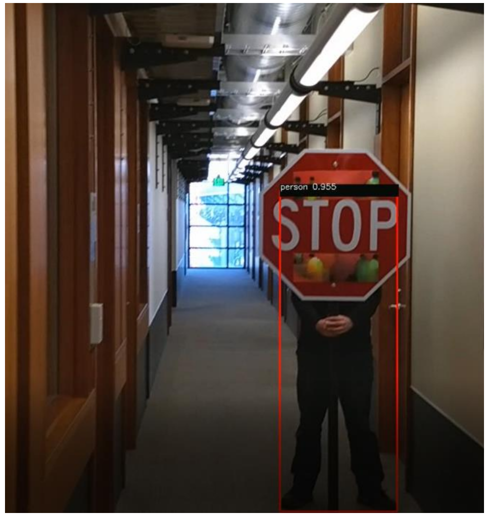
Add to Your Schedule

Andrew Ilyas (Massachusetts Institute of Technology), Logan Engstrom (Massachusetts Institute of Technology), Anish Athalye (Massachusetts Institute of Technology)

- Adversarial eye glasses

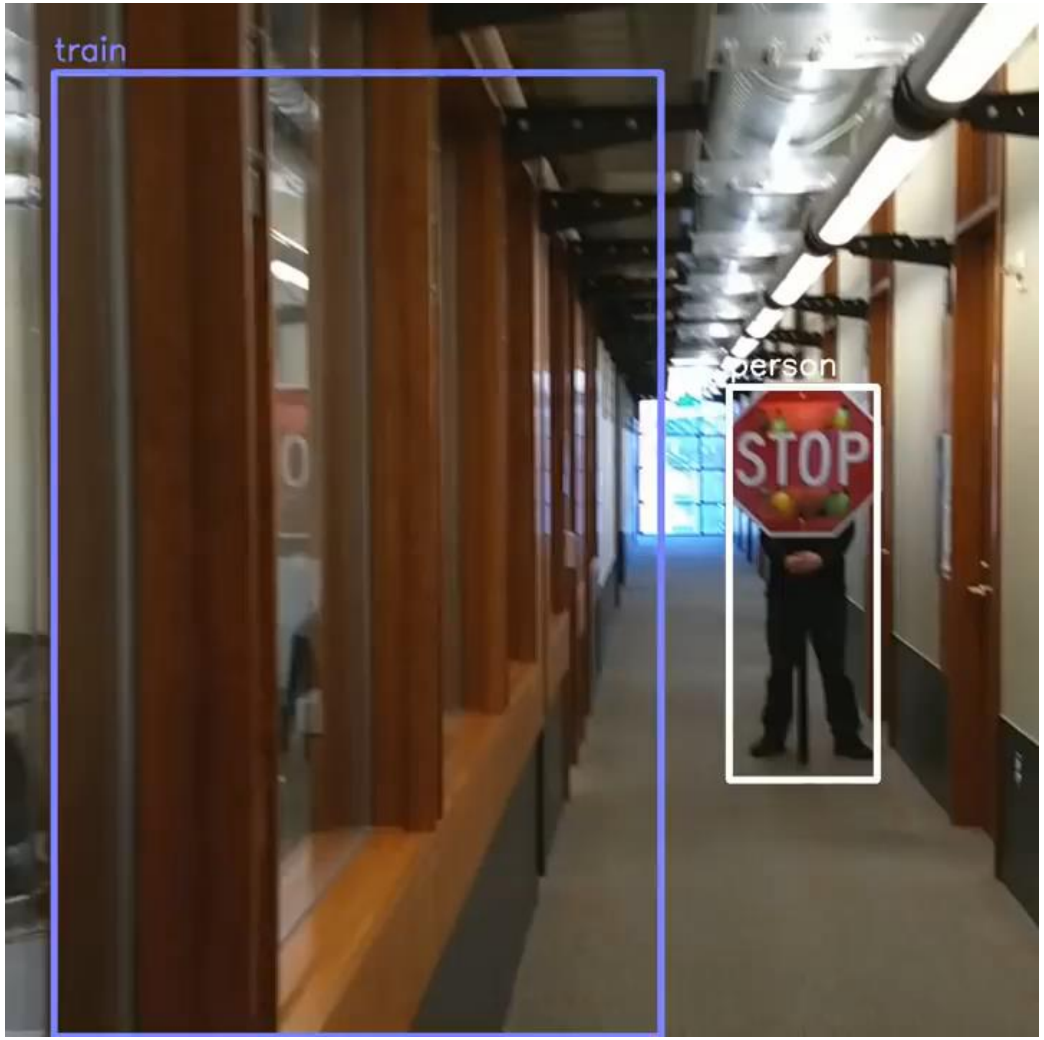
Adversarial examples in physical world (1)

- Real-time traffic sign detector



Robust Physical-World Attacks on Deep Learning Visual Classification

Kevin Eykholt*[1], Ivan Evtimov*[2], Earlence Fernandes[2], Bo Li[3],
Amir Rahmati[4], Chaowei Xiao[1], Atul Prakash[1], Tadayoshi Kohno[2], and Dawn Song[3]

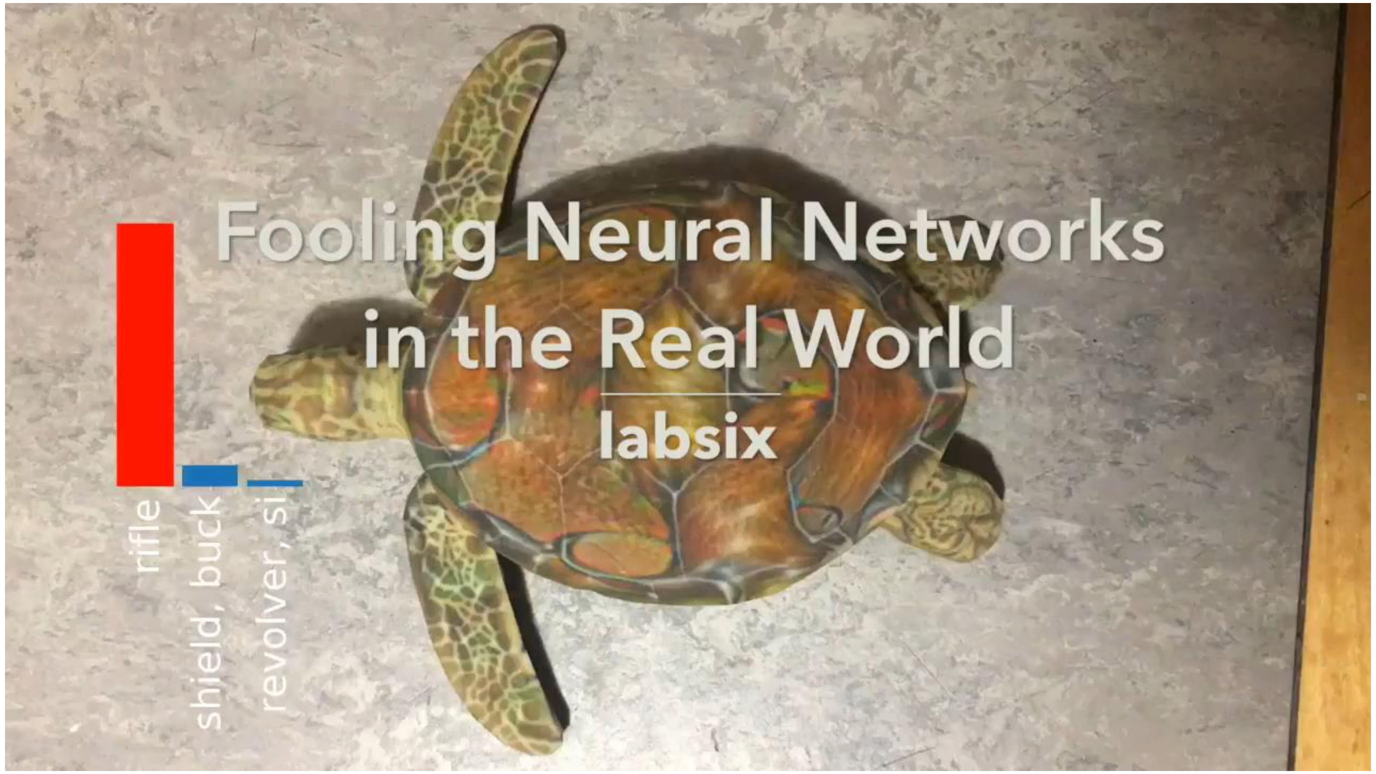


Adversarial examples in physical world (2)

- 3D-printed adversarial turtle

Synthesizing Robust Adversarial Examples

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



Adversarial examples in physical world (3)

- Adversarial eye glasses that fool face detector

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

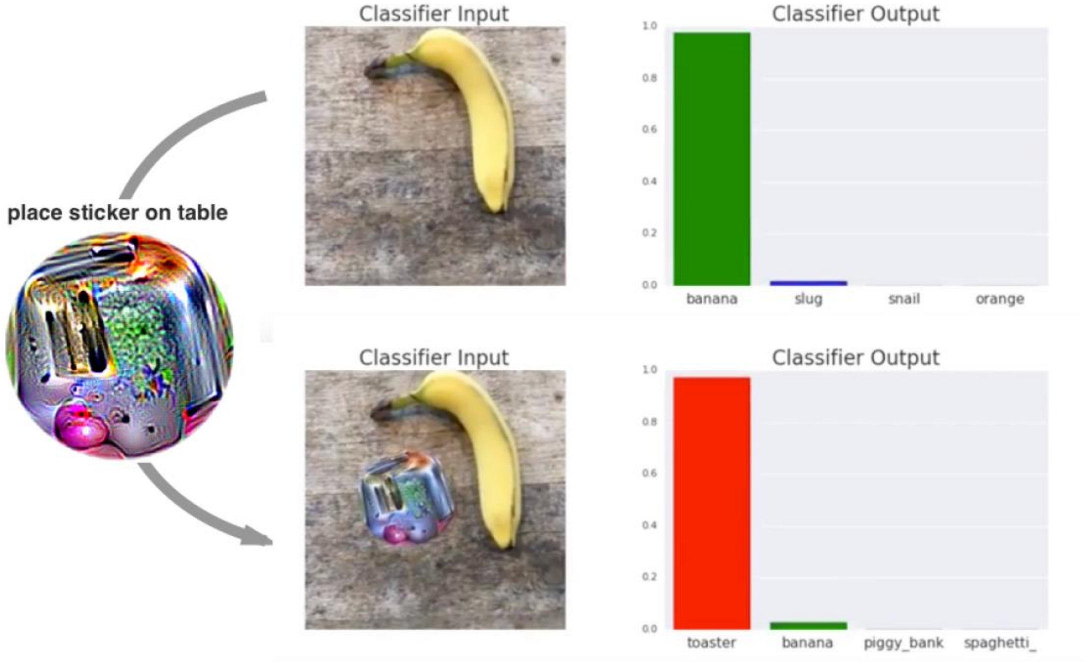
Mahmood Sharif
Carnegie Mellon University
Pittsburgh, PA, USA
mahmoods@cmu.edu

Sruti Bhagavatula
Carnegie Mellon University
Pittsburgh, PA, USA
srutib@cmu.edu

Lujo Bauer
Carnegie Mellon University
Pittsburgh, PA, USA
lbauer@cmu.edu

Michael K. Reiter
University of North Carolina
Chapel Hill, NC, USA
reiter@cs.unc.edu

- Adversarial sticker

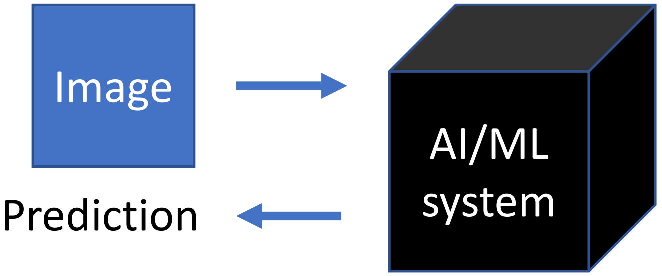


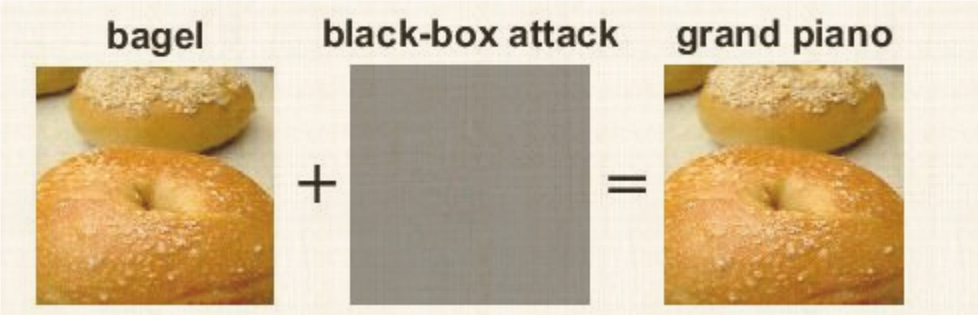
Adversarial Patch

Tom B. Brown, Dandelion Mané*, Aurko Roy, Martín Abadi, Justin Gilmer
{tombrown,dandelion,aurkor,abadi,gilmer}@google.com

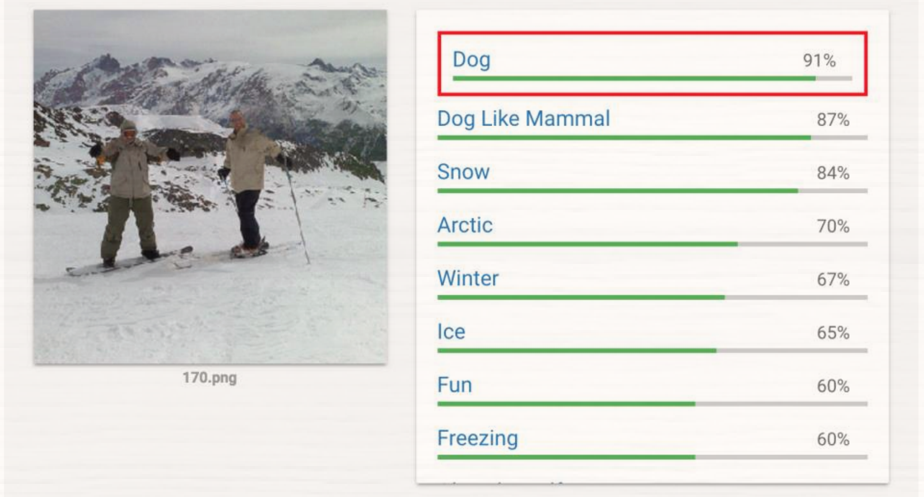
Adversarial examples in black-box models

- **White-box setting**: adversary knows everything about your model
- **Black-box setting:** craft adversarial examples with limited knowledge about the target model
    - Unknown training procedure/data/model
    - Unknown output classes
    - Unknown model confidence





Black-box attack via iterative model query (ZOO)



Targeted black-box attack on Google Cloud Vision

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 Adversarial Attacks with Limited Queries and Information, Andrew Ilyas*, Logan Engstrom*, Anish Athalye*, and Jessy Lin*, ICML 2018
Source: https://www.labsix.org/partial-information-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…

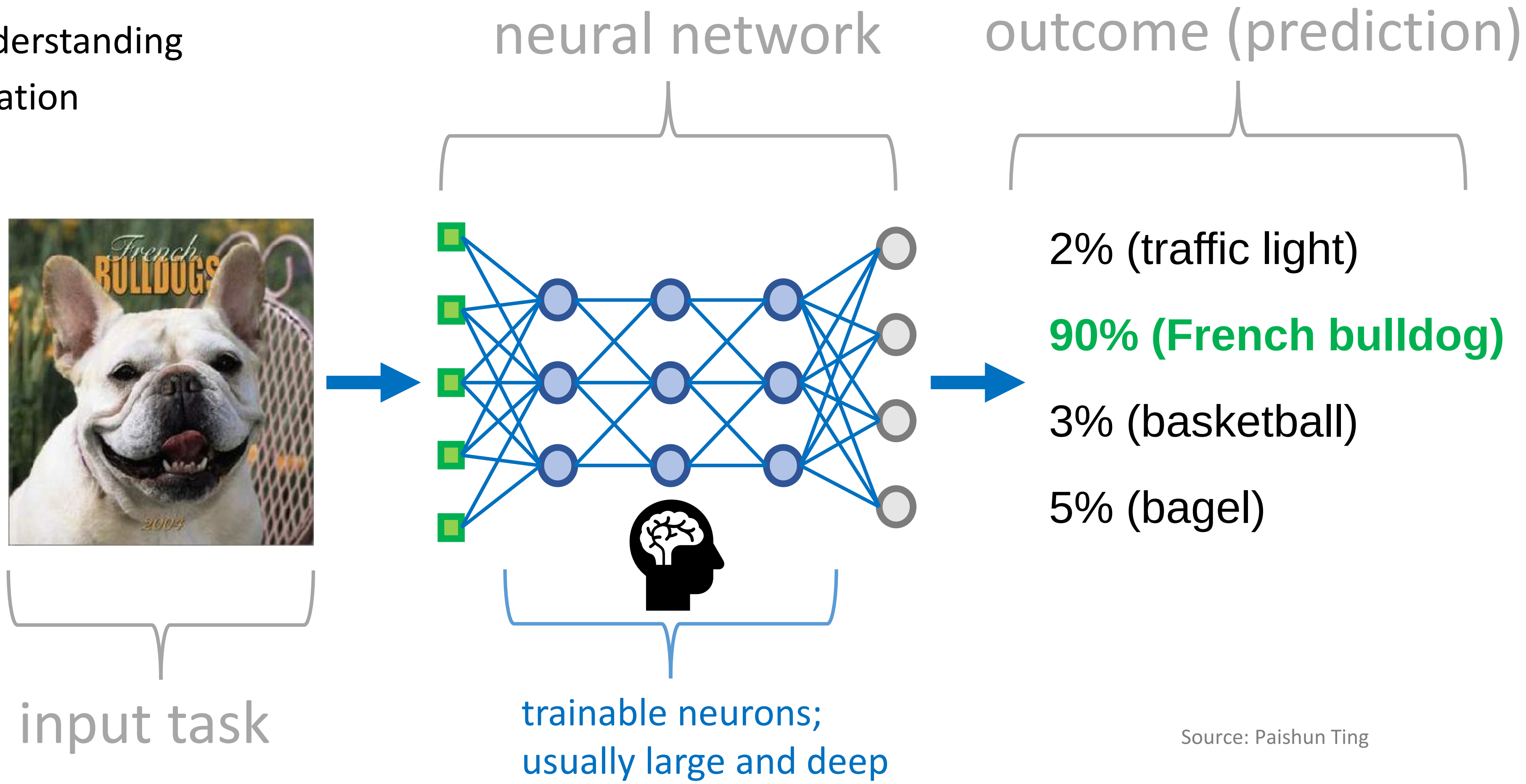
Source: Paishun Ting

Where do adversarial examples come from?

- What is the common theme of adversarial examples in different domains?

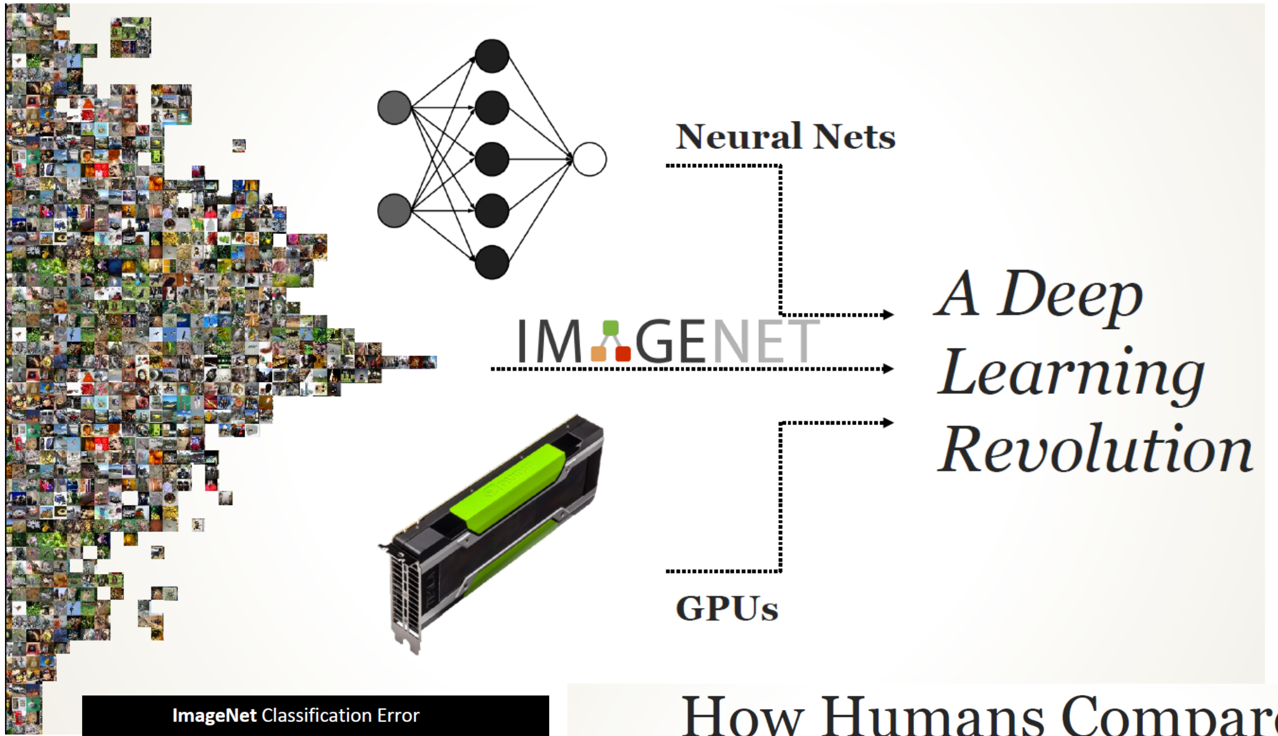
Neural Networks: The Engine for Deep Learning

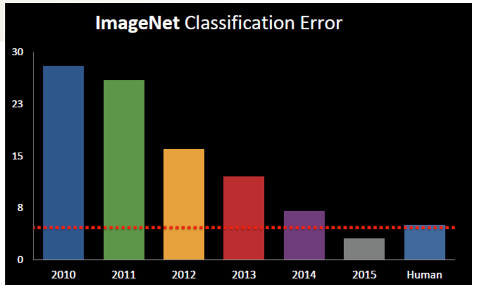
- Applications of neural networks
- ❑ Image processing and understanding
- ❑ Object detection/classification
- ❑ Chatbot, Q&A
- ❑ Machine translation
- ❑ Speech recognition
- ❑ Game playing
- ❑ Robotics
- ❑ Bioinformatics
- ❑ Creativity
- ❑ Drug discovery
- ❑ Reasoning
- ❑ And still a long list…



Source: Paishun Ting

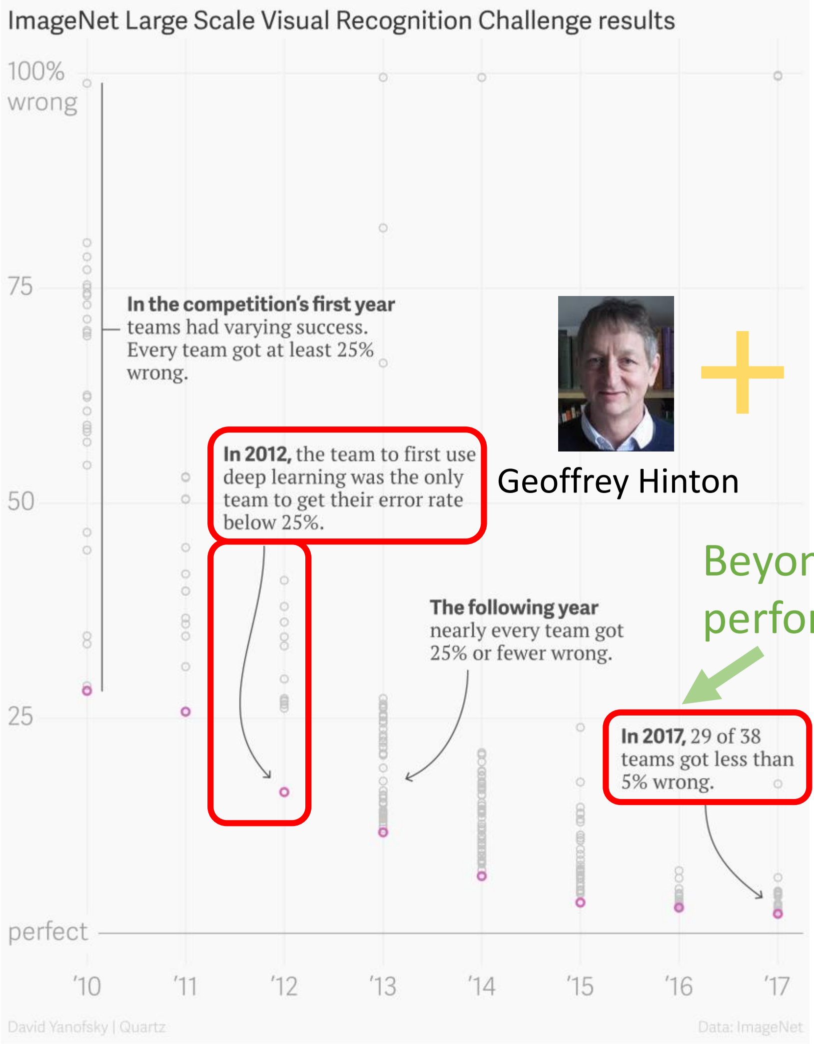
The ImageNet Accuracy Revolution and Arms Race



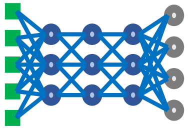


How Humans Compare

Human	GoogLeNet
5.1% Top-5 error rate	**6.8%** Top-5 error rate



Geoffrey Hinton



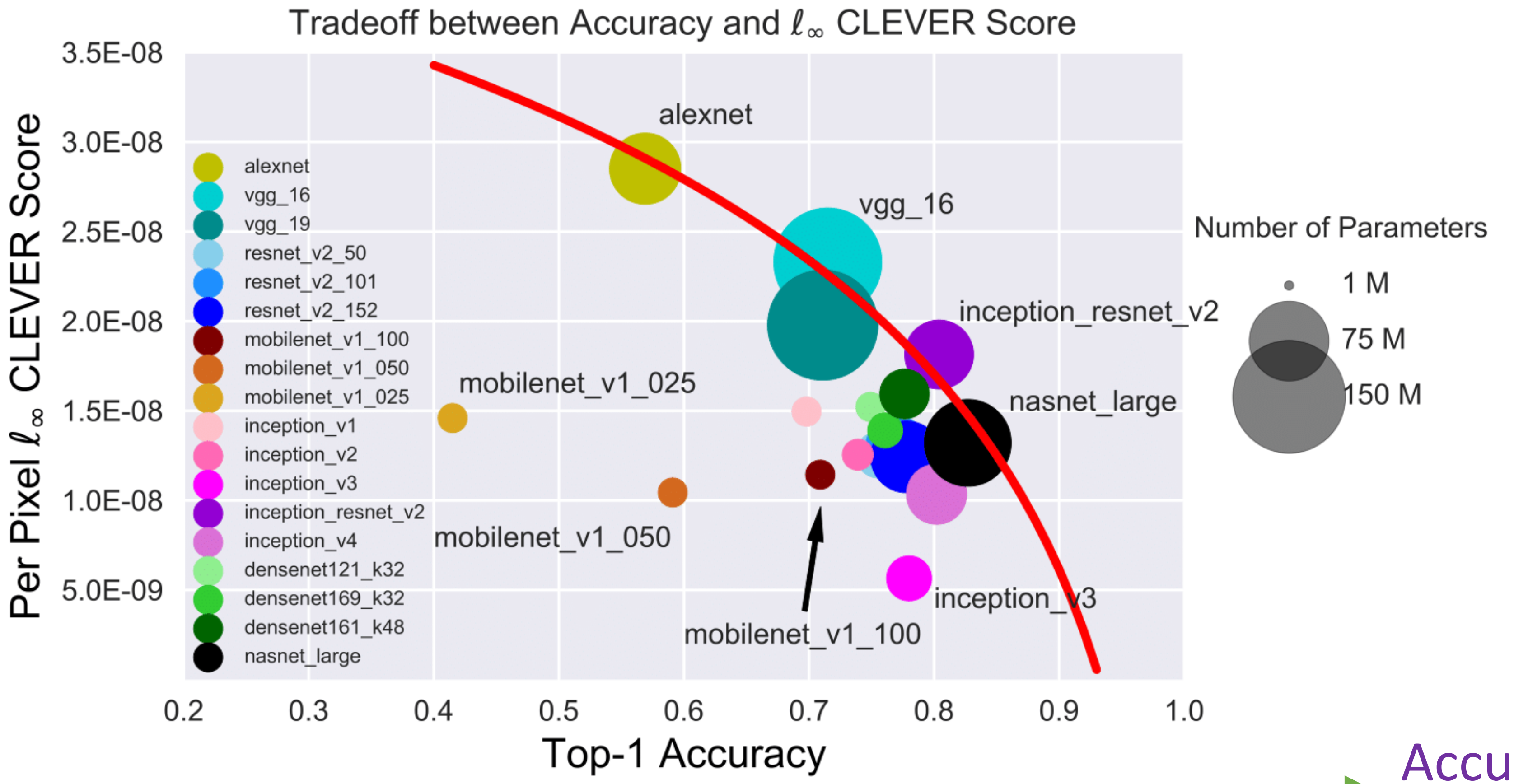
Beyond human performance

What's Next?

Source: http://image-net.org/challenges/talks_2017/imagenet_ilsvrc2017_v1.0.pdf
Source: https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world/

Accuracy ≠ Adversarial Robustness

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



Robustness

Accuracy

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

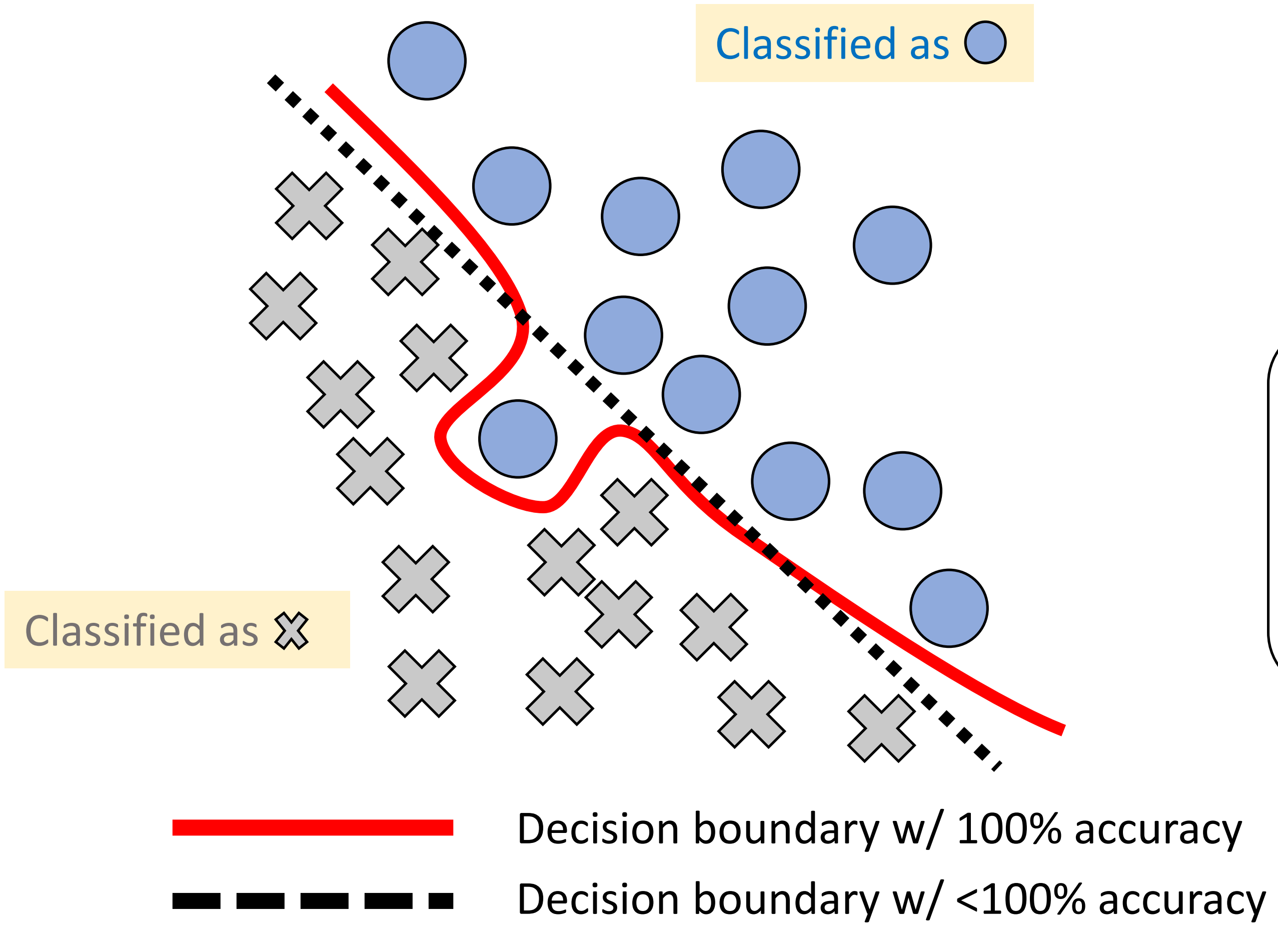
Is Robustness the Cost of Accuracy? A Comprehensive Study on the Robustness of 18 Deep Image Classification Models, Dong Su*, Huan Zhang*, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao, ECCV 2018

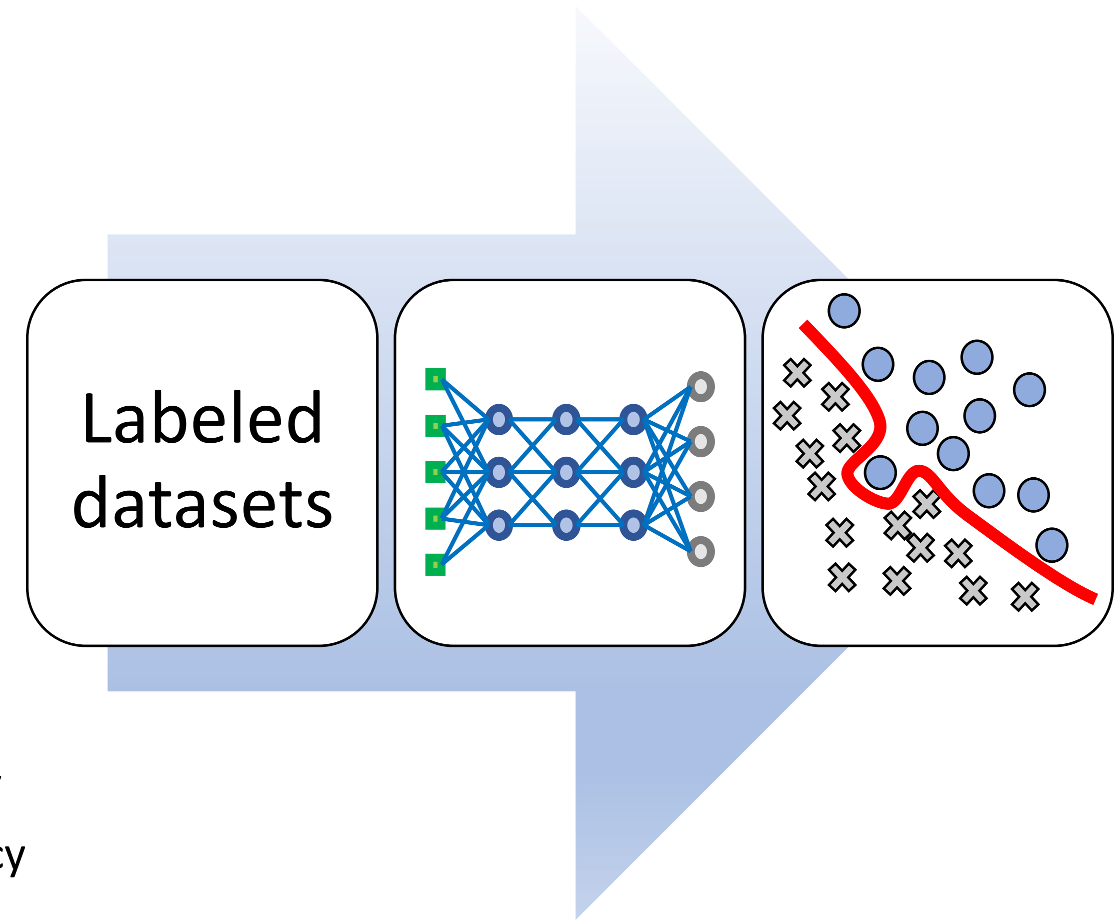
How can we <u>measure</u> and improve adversarial robustness of my AI/ML model?

An explanation of origins of adversarial examples

The CLRVER score for robustness evaluation

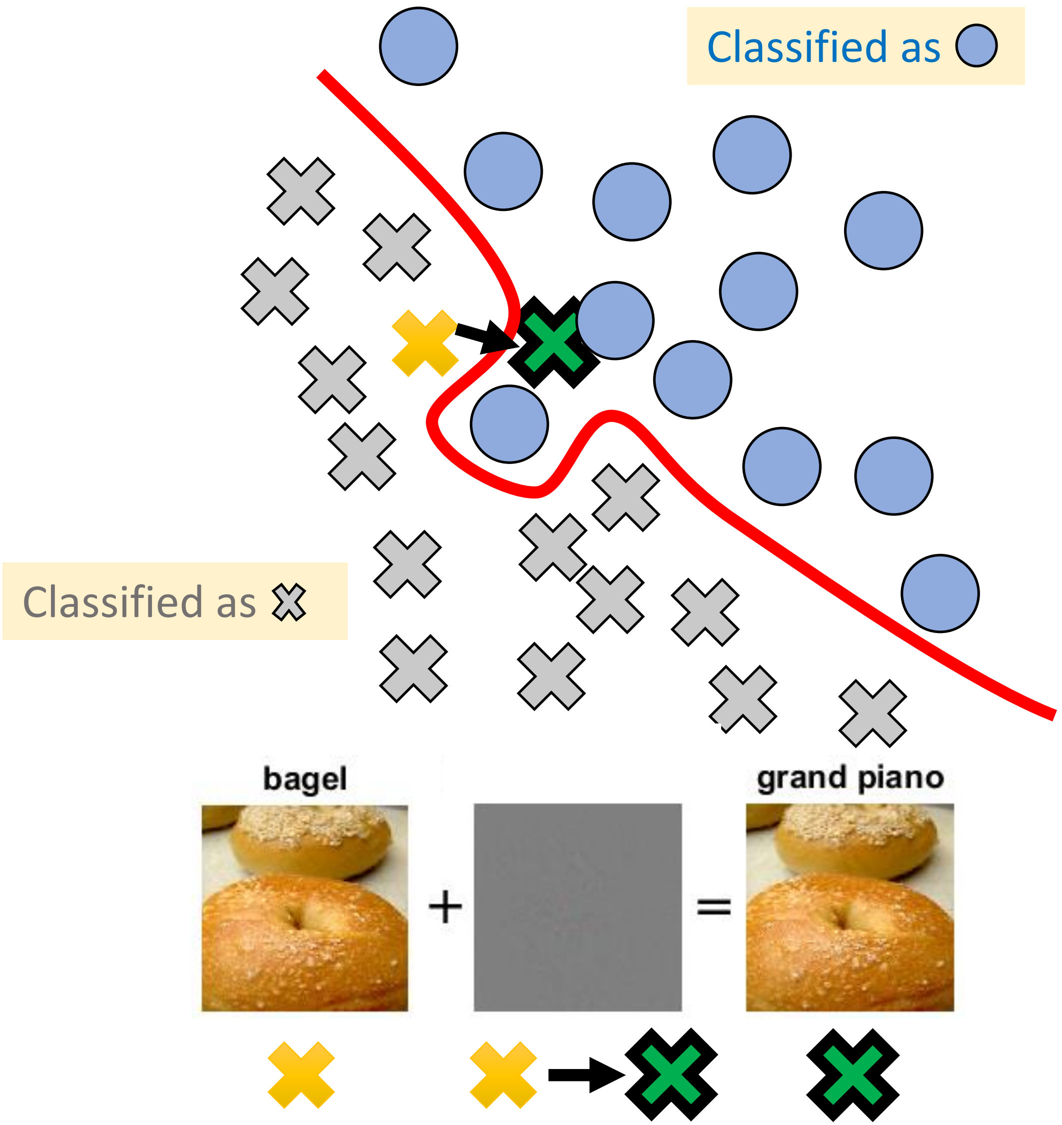
Learning to classify is all about drawing a line

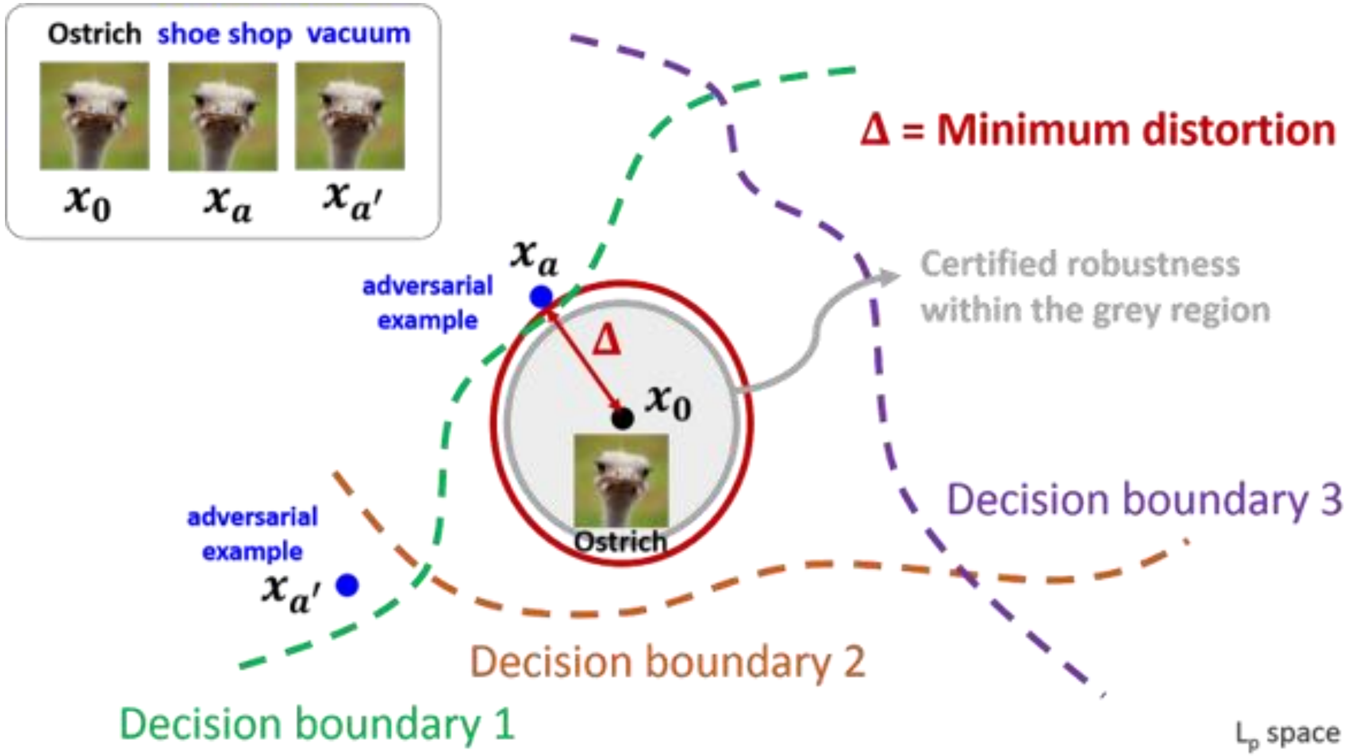




Source: Paishun Ting

Connecting adversarial examples to model robustness



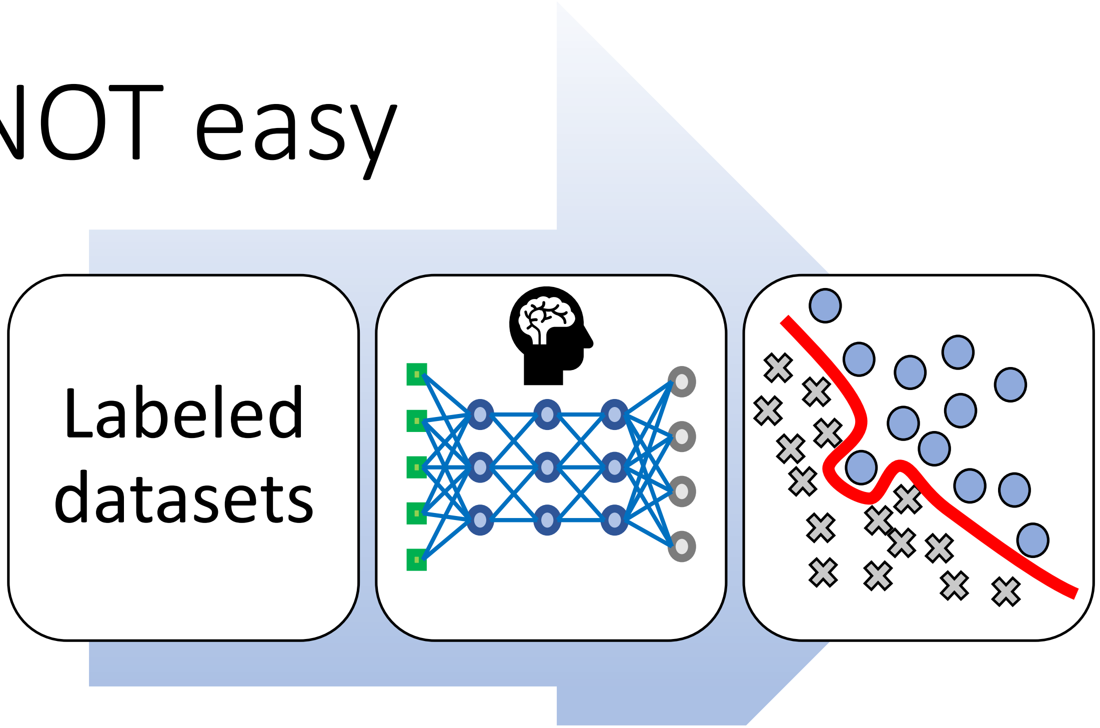


- Robustness evaluation: how close a refence input is to the (closest) decision boundary

Source:Paishun-Ting, Tsui-Wei Weng

Robustness evaluation 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]

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

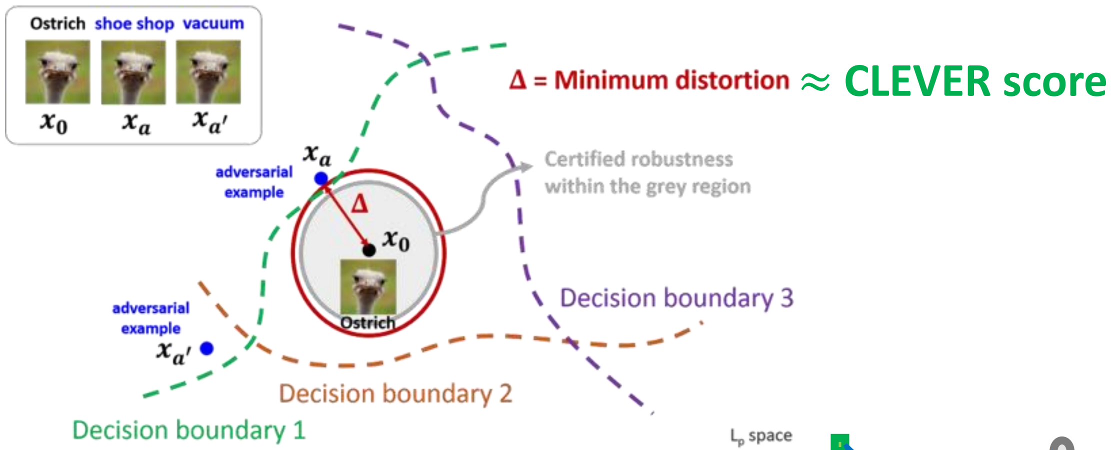
❑Can provide a robustness certificate for safety-critical or reliability-sensitive applications: e.g., no attacks can alter the decision of the AI model if the attack strength is limited

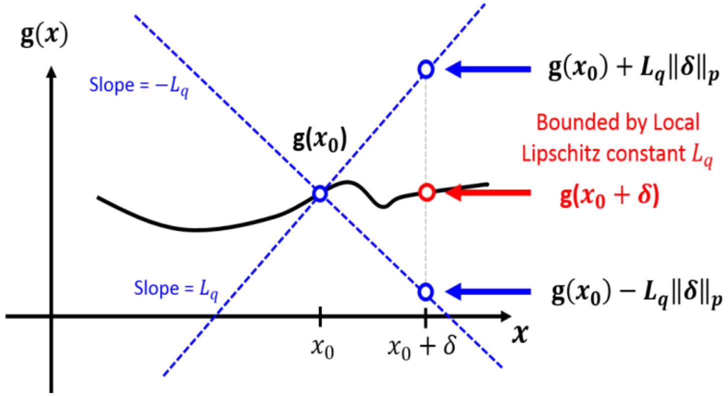
Optimal verification is provably difficult for large neural nets – computationally impractical

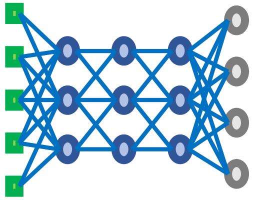
- Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks, Guy Katz, Clark Barrett, David Dill, Kyle Julian, Mykel Kochenderfer, CAV 2017
- Efficient Neural Network Robustness Certification with General Activation Functions, Huan Zhang*, Tsui-Wei Weng*, Pin-Yu Chen, Cho-Jui Hsieh, and Luca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

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







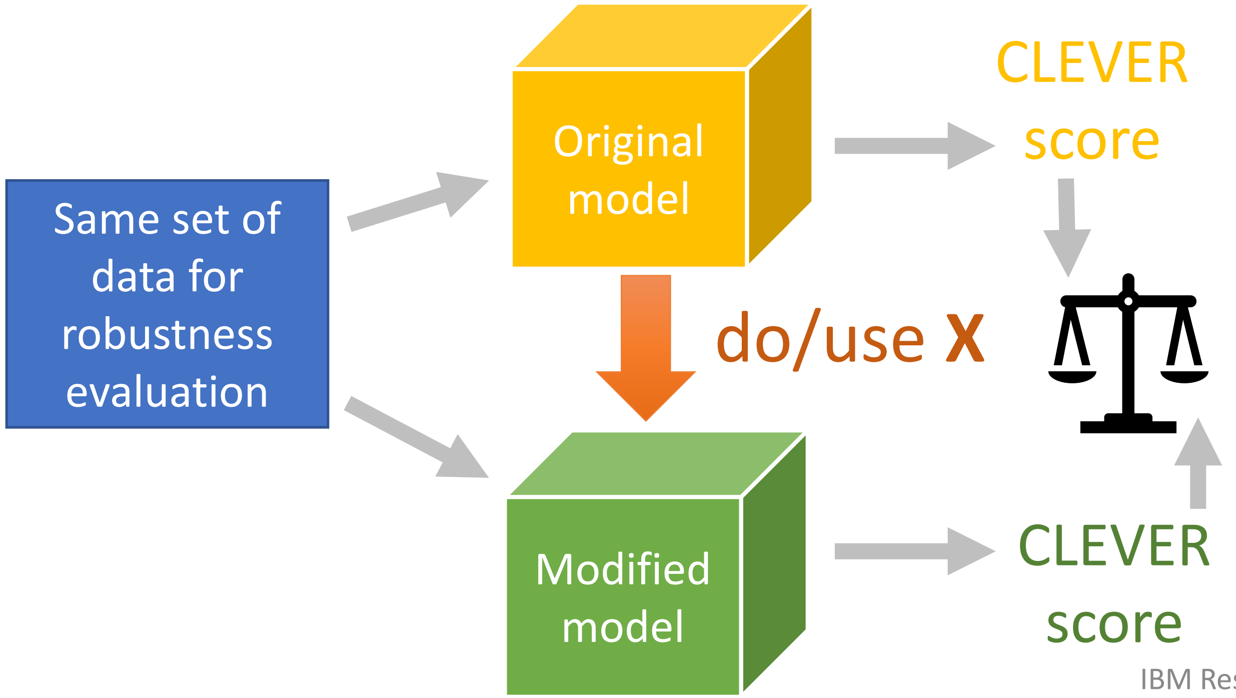
input-output perturbation analysis of neural net

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

How do we use CLEVER?

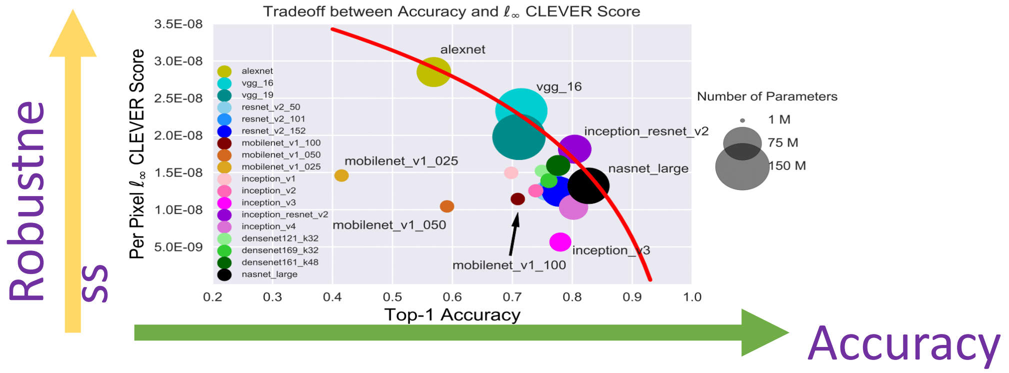
Before-After robustness comparison

- 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



IBM Research AI

Examples of CLEVER

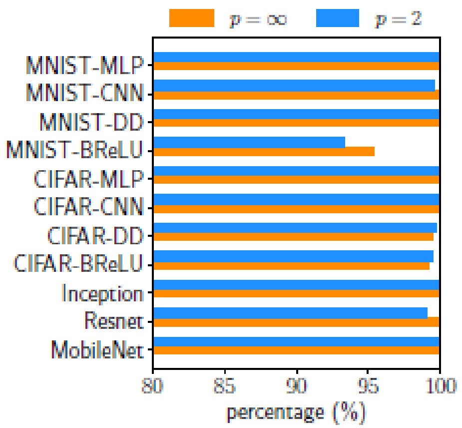
- CLEVER enables robustness comparison between <u>different</u>

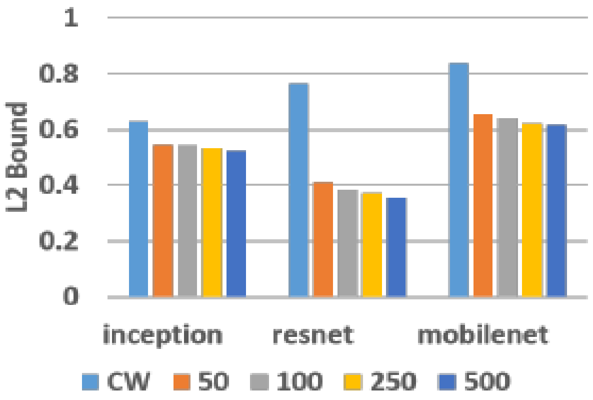
❑Threat models

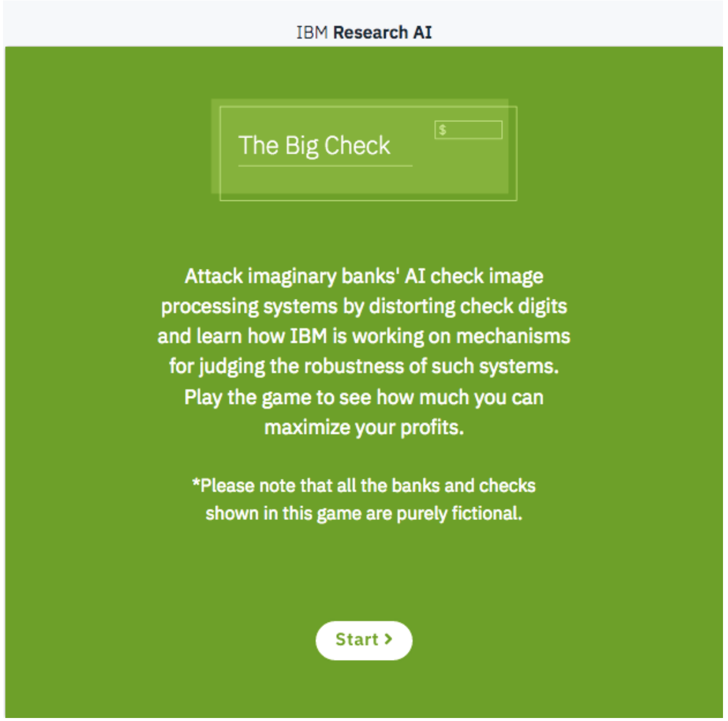
❑Datasets

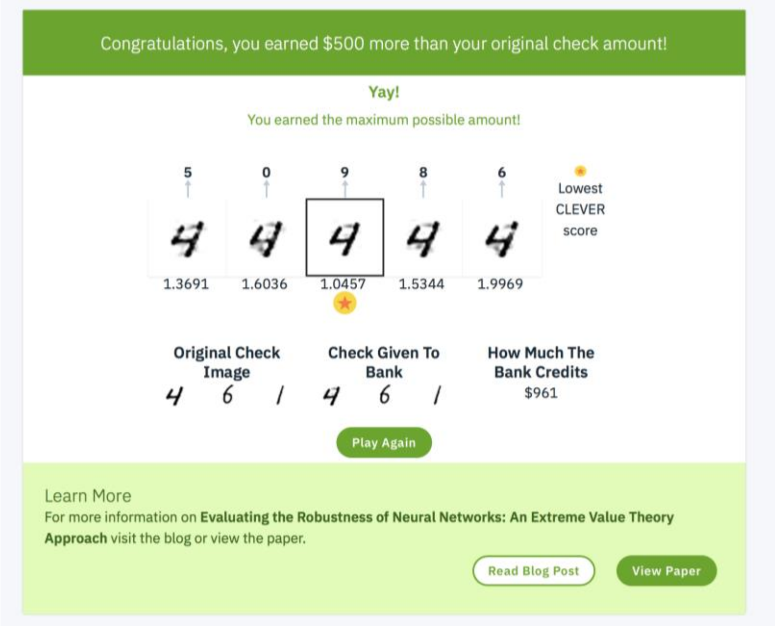
❑Neural network architectures

❑Defense mechanisms









Where to Find CLEVER? It's 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 →

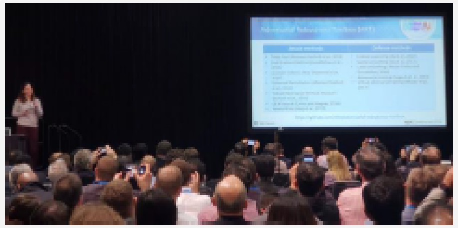
Load classifier model (Keras, TF, PyTorch etc) →

Perform attack →

Evaluate robustness →

```python
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

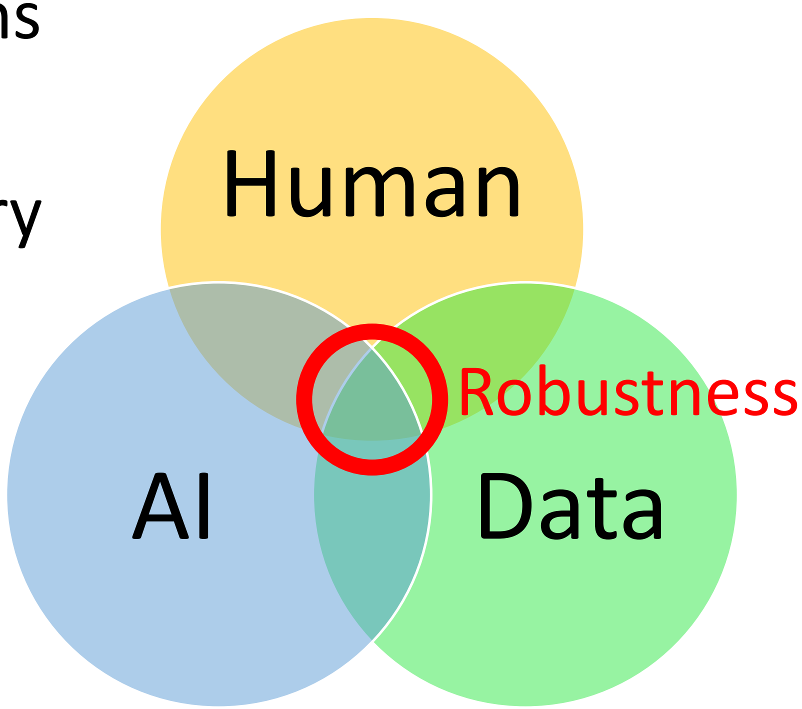
IBM Adversarial Robustness Toolbox beschermt tegen kwaadaardige AI

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

Also available at https://github.com/IBM/CLEVER-Robustness-Score

Take-aways

- Adversarial robustness is a new AI standard
  - ❑Robustness does not come for free: adversarial examples exist in digital space, physical world, and different domains
  - ❑High accuracy ≠ Good robustness
  - ❑Arms race: adversary-aware AI v.s. AI for adversary
- How to evaluate the robustness of my AI model?
  - ❑CLEVER: an attack-independent robustness score
  - ❑Robustness comparison in before-after setting
  - ❑Where to find CLEVER? It’s ART!



Beyond Robustness: Trusted AI

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

Acknowledgement

- Collaborators: Tsui-Wei Weng(MIT), Luca Daniel(MIT), Honnge Chen(MIT) Huan Zhang(UCLA), Cho-Jui Hsieh(UCLA), Jinfeng Yi(JD AI), Yupeng Gao(IBM), Bhanukiran Vinzamuri(IBM), Sijia Liu(IBM), Yash Sharma, Su Dong, Chun-Chen Tu(UMich), Paishun Ting(Umich)
- MIT-IBM Watson AI Lab: David Cox, Lisa Amini
- IBM Research AI – Learning Group: Payel Das, Saska Mojsilovic
- IBM AI-Security Group: Ian Molloy, Mathieu Sinn, and their teams
- IBM Big Check Demo: Casey Dugan and her team
- IBM DLaaS Group: Evelyn Duesterwald and her team

❑ Personal Website: www.pinyuchen.com

❑ Twitter: pinyuchen.tw