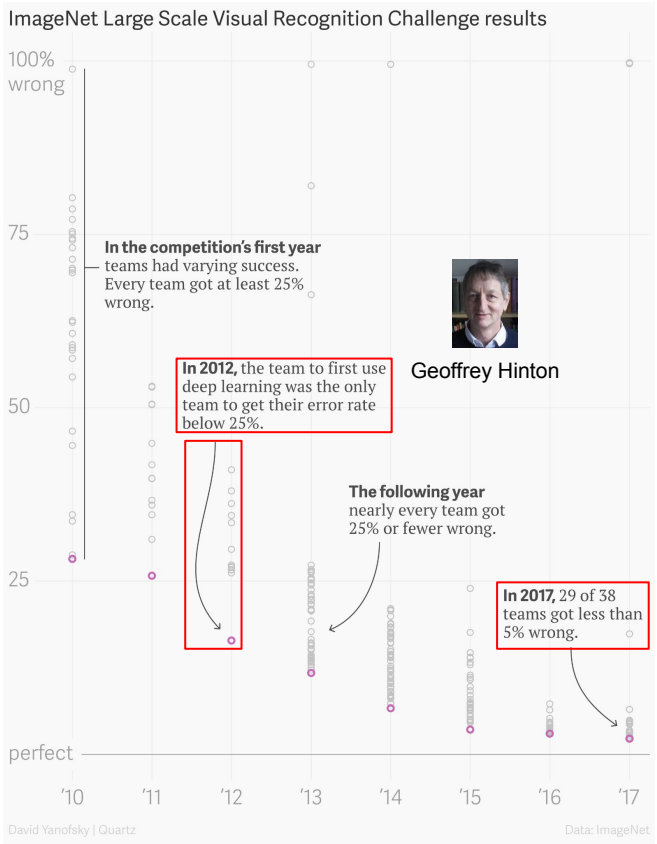
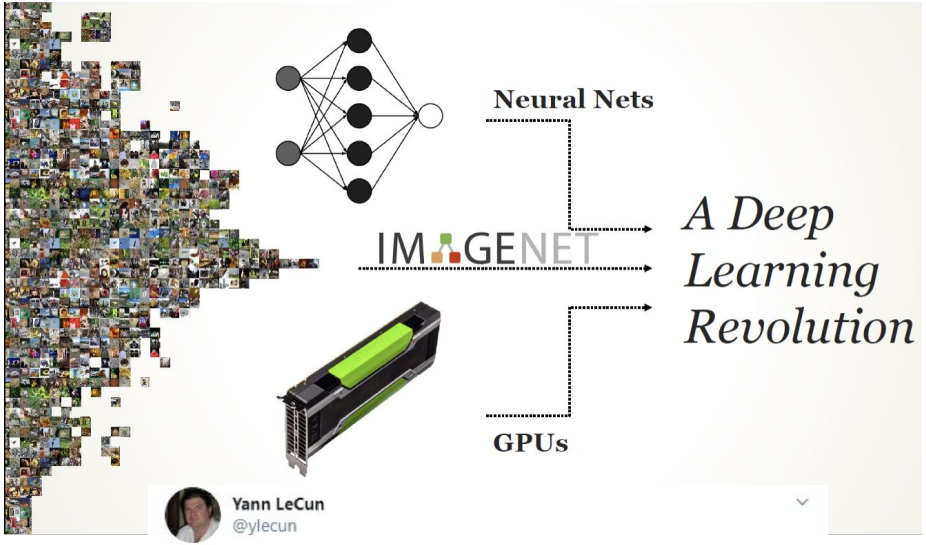
The Deep Learning Revolution. What's next?



Geoffrey Hinton

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

What's next?



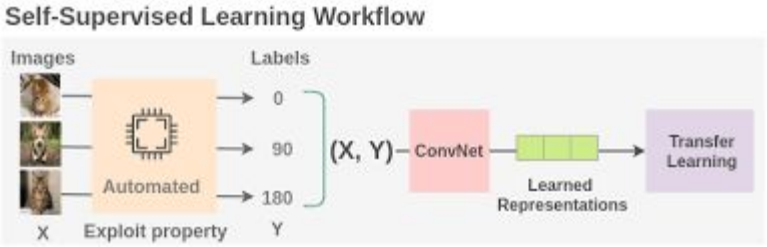
Yann LeCun
@ylecun

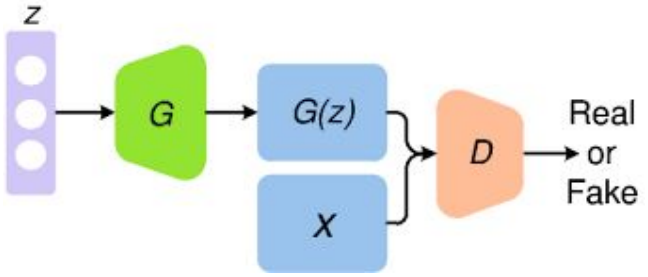
Replying to @ylecun @GaryMarcus and @titudeadjust

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

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

Prompt > Gradient descent is a first-order iterativ
Prompt > Artificial intelligence (AI), sometimes ca
Prompt > ZDNet is a business technology news websit
Prompt > OpenAI is an artificial intelligence resea
ZDNet > GPT-3 is the next word in AI
Prompt > Deep learning (also known as deep structut
Prompt > Unsupervised learning is a type of machine
Prompt > Labeled data is a group of samples that ha
Prompt > Conditional probability is a neasure of th





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

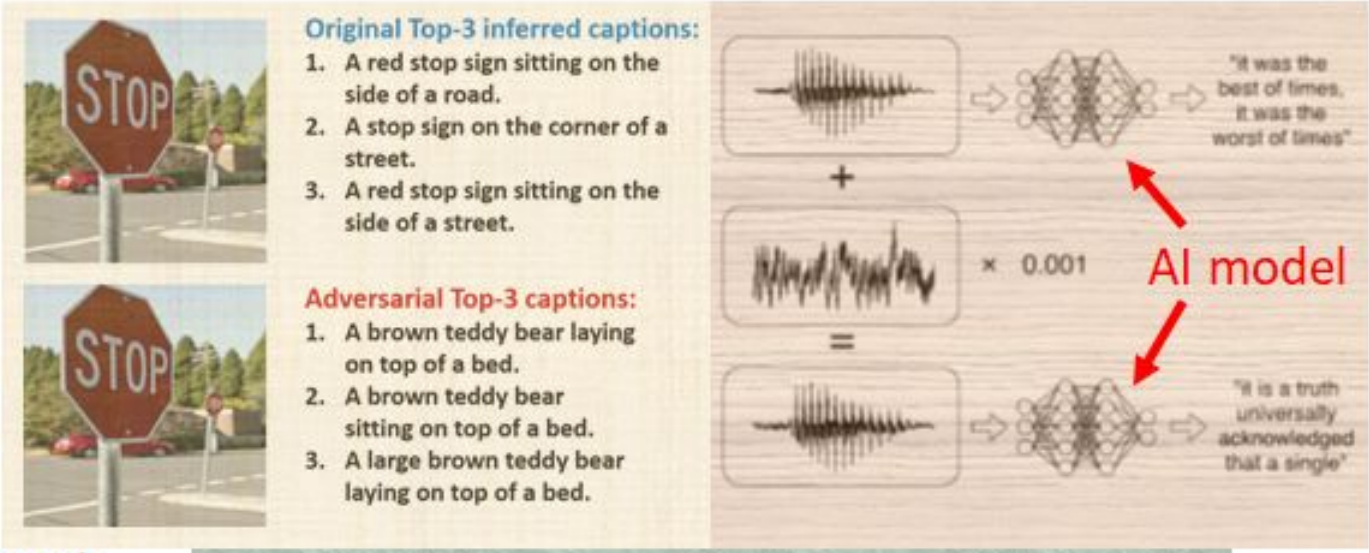
What is wrong with this AI model?

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

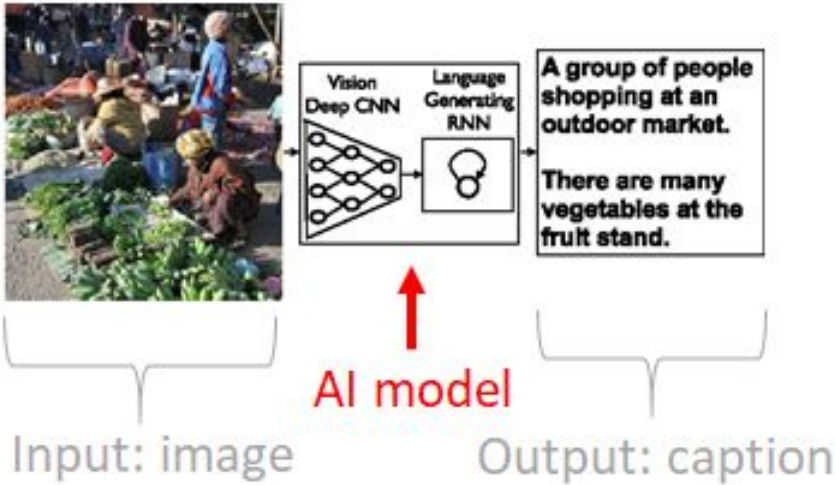
EAD: Elastic Net Attacks to Deep Neural Networks via Adversarial Examples, P. Y. Chen*, Y. Sharma*, H. Zhang, J. Yi, and C. J. Hsieh, AAAI 2018

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



Original Top-3 inferred captions:

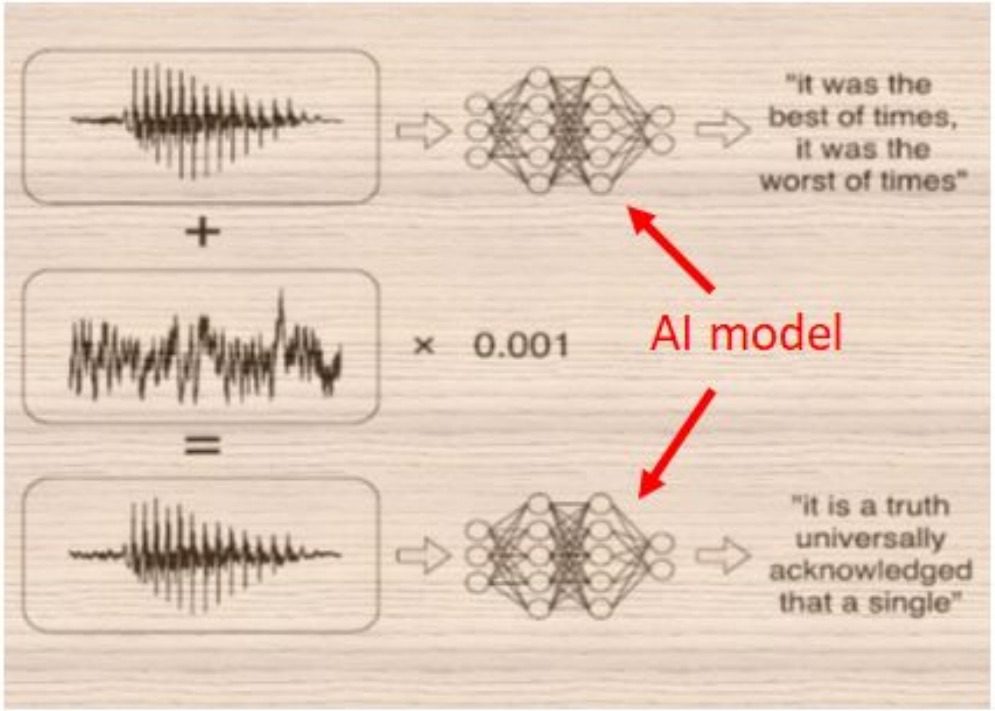
1. **A red stop sign sitting on the side of a road.**
2. **A stop sign on the corner of a street.**
3. **A red stop sign sitting on the side of a street.**

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, Alexander Toshev, 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



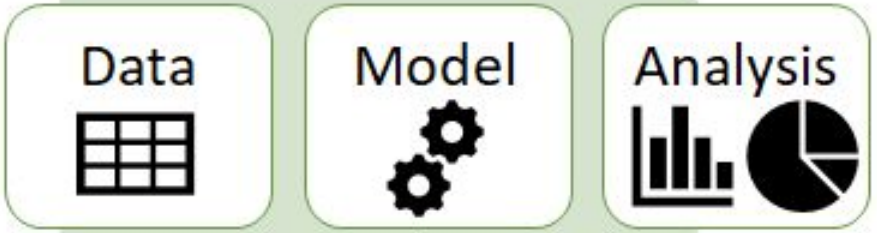
without the dataset the article is useless

What did your hear?

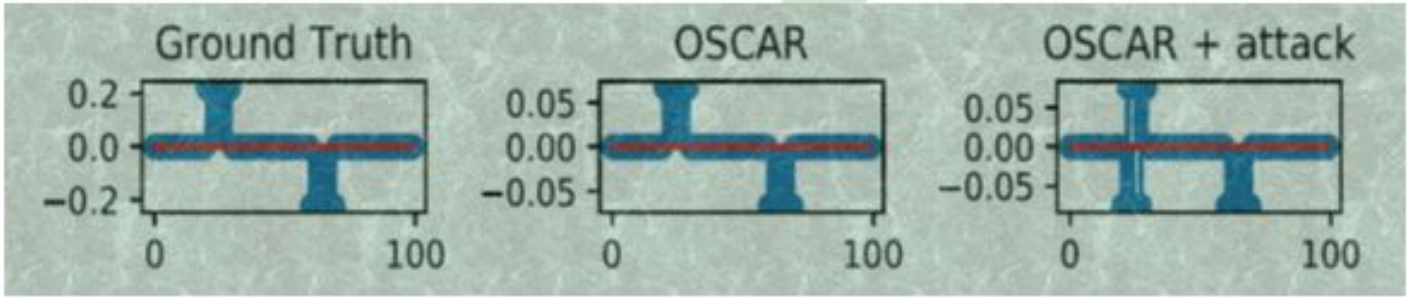
okay google browse to evil.com

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**

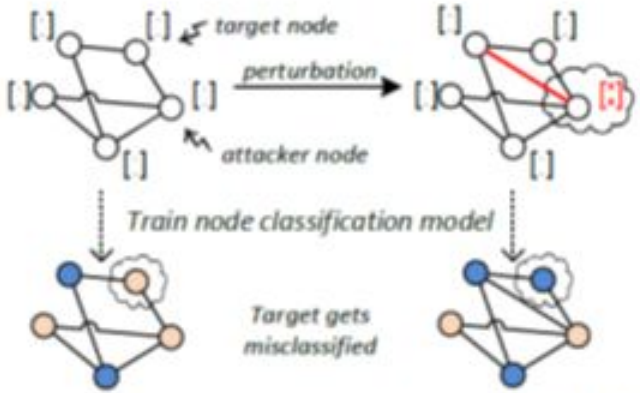
- Targeted phrase attack for text summarization. Target: "police arrest"

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.

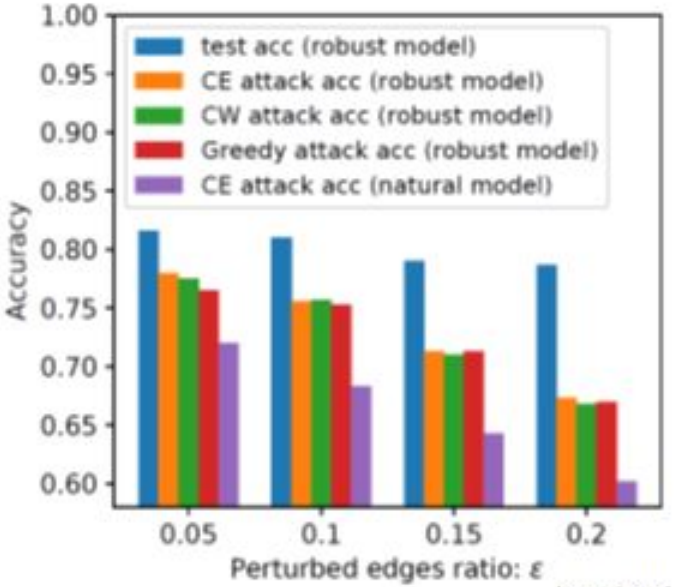
Minhao Cheng, Jinfeng Yi, Pin-Yu Chen, Huan Zhang, and Cho-Jui Hsieh, "Seq2Sick: Evaluating the Robustness of Sequence-to-Sequence Models with Adversarial Examples," *AAAI Conference on Artificial Intelligence (AAAI)*, 2020

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.

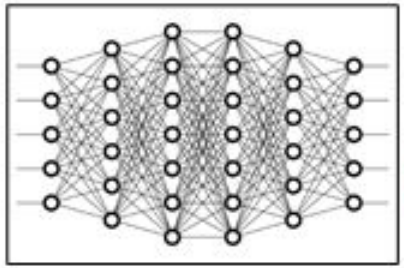
Adversarial examples in deep reinforcement learning

- Observation (state) perturbation for policy/reward degradation

Sequential Inputs

Frame under Attack

Deep Reinforcement Learning Agent

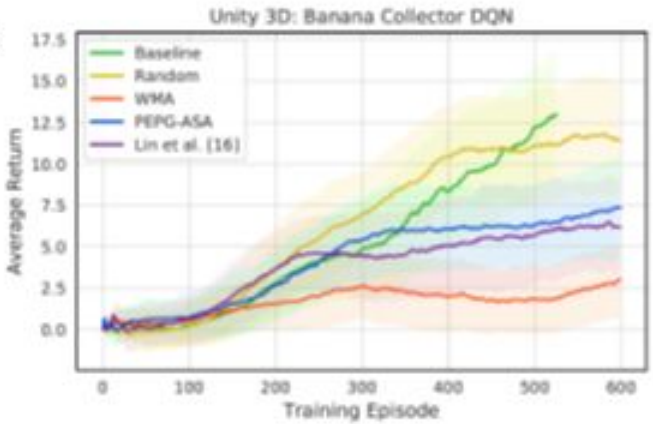


Output Actions

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

Output Action at time = t

"Left"



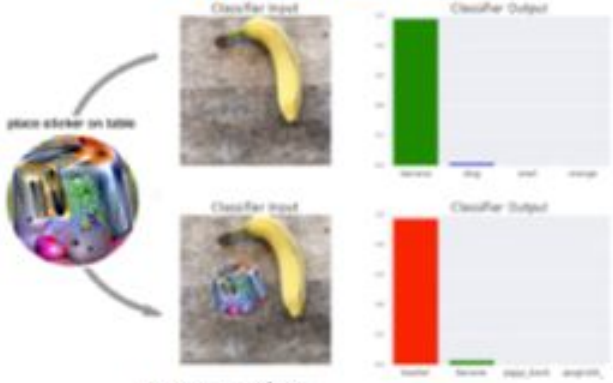
Credit: Chao-Han Huck Yang@GIT

Adversarial examples in physical world

- Real-time traffic sign detector

- 3D-printed adversarial turtle

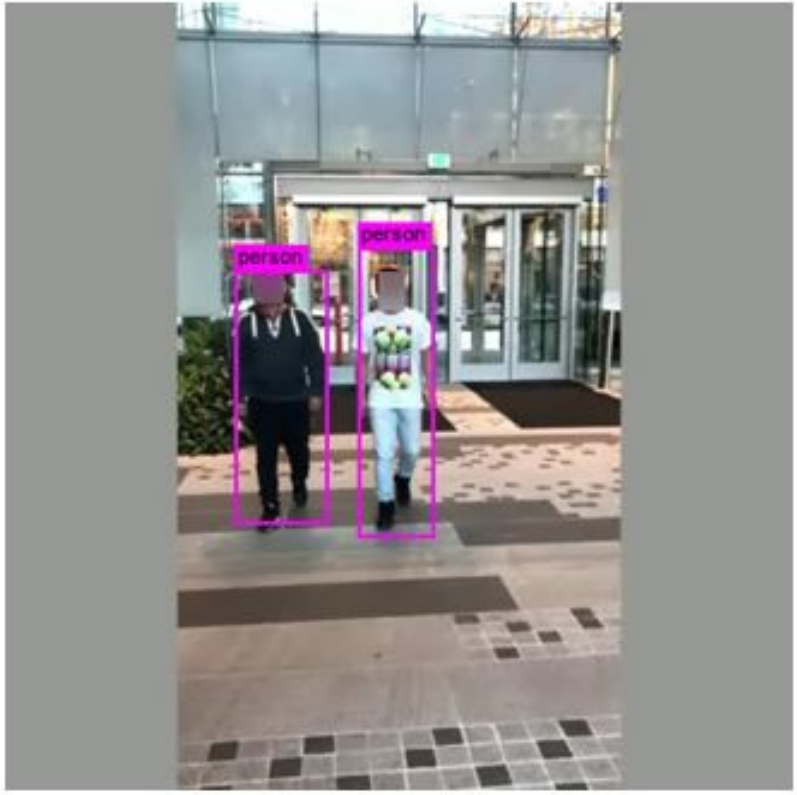
- Adversarial patch



- Adversarial eye glasses

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

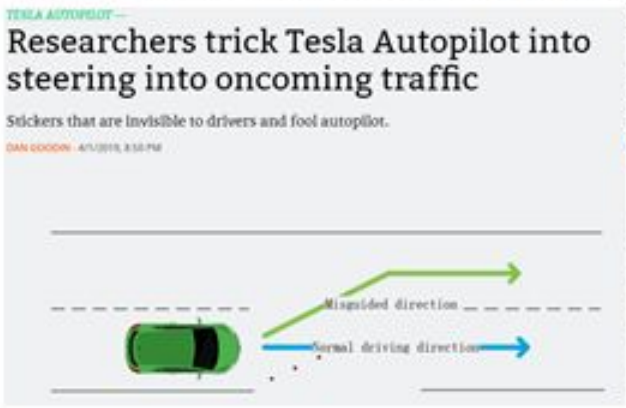
Adversarial T-shirt! Evading Person Detectors in A Physical World. Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang, Xue Lin, ECCV 2020

2. Adversarial AI

Why adversarial (worst-case) robustness matters?

- Prediction-evasive manipulation on a deployed AI model

1. Build trust in AI: address inconsistent perception and decision making between humans and machines & misinformation
2. Assess negative impacts in high-stakes, safety-critical tasks
3. Understand limitation in current machine learning methods
4. Prevent loss in revenue and reputation
5. Ensure safe and responsible use in AI



AI technology: Jewel of the Crown

Adversarial ML Threat Matrix

https://github.com/mitre/advmlthreatmatrix

AI Incidence Database

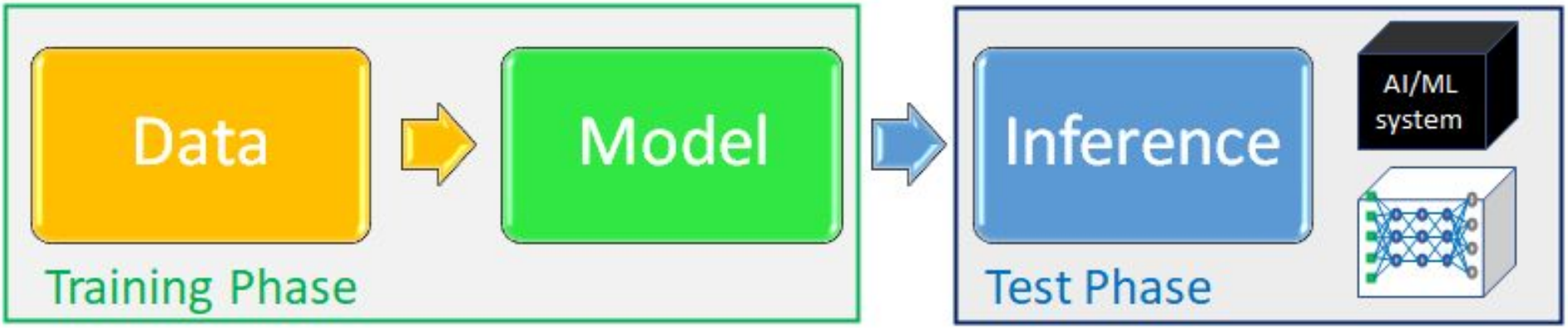
https://incidentdatabase.ai

- 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/

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