The Promise & Peril of Deep Learning for Cybersecurity

Dr. John McKay

Research & Development Engineer

Applied Research Laboratory

Pennsylvania State University

PennState Applied Research Laboratory

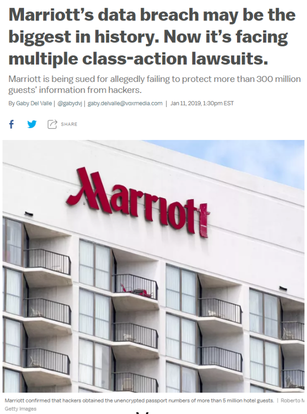
March 28, 2018

5th Cybersecurity Workshop

Table of Contents

Outline

Independent


Marriott's data breach may be the biggest in history. Now it's facing multiple class-action lawsuits.

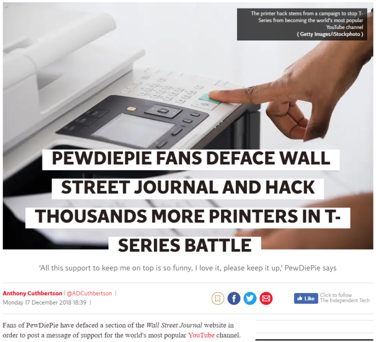
Marriott is being sued for allegedly failing to protect more than 300 million guests' information from hackers.

By Gaby Del Valle | @gabydvj | gaby.delvalle@voxmedia.com | Jan 11, 2019, 1:30pm EST

SHARE

Marriott confirmed that hackers obtained the unencrypted passport numbers of more than 5 million hotel guests. | Roberto M Getty Images

Vox



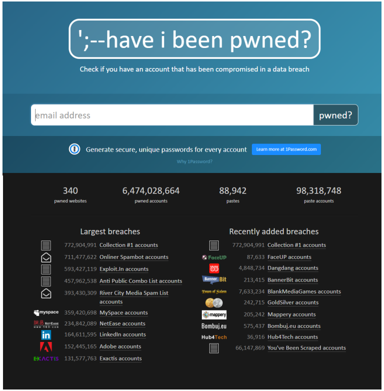
PEWDIEPIE FANS DEFACE WALL STREET JOURNAL AND HACK THOUSANDS MORE PRINTERS IN T-SERIES BATTLE

'All this support to keep me on top is so funny, I love it, please keep it up,' PewDiePie says

Anthony Cuthbertson | @ADCuthbertson |
Monday 17 December 2018 18:39 |

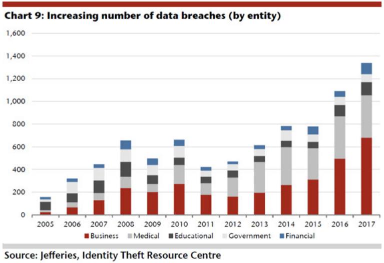
Like Click to follow The Independent Tech

Fans of PewDiePie have defaced a section of the *Wall Street Journal* website in order to post a message of support for the world's most popular YouTube channel.

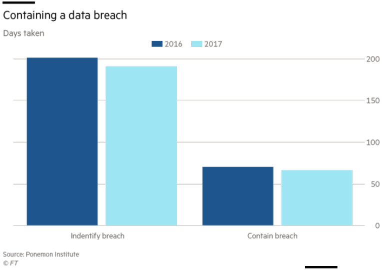


haveibeenpwned.com

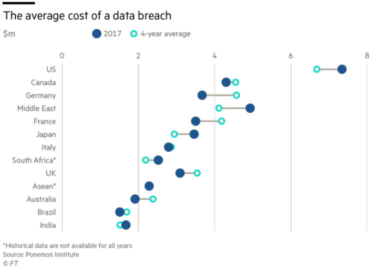
Cybersecurity



marketwatch.com



Financial Times



- As threats have grown and hackers have developed increasingly sophisticated strategies for accessing sensitive data, commercial/government/etc. organizations have started looking to **deep learning** to aid in prevention and detection.
- This talk is designed to cover why deep learning is an attractive avenue and why we should be careful of its hidden flaws.

Can AI Become Our New Cybersecurity Sheriff?

Naveen Joshi Contributor
COGNITIVE WORLD Contributor Group
AI & Big Data

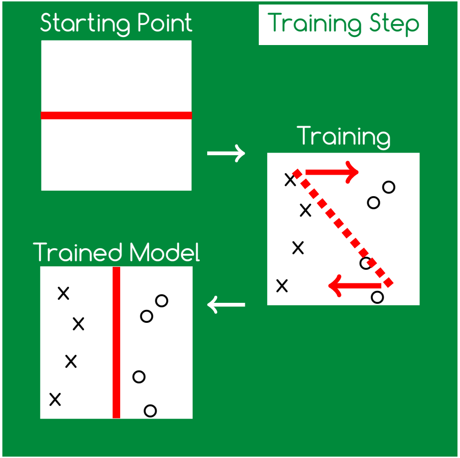
AI, the new sheriff DEPOSITPHOTOS ENHANCED BY COGWORLD

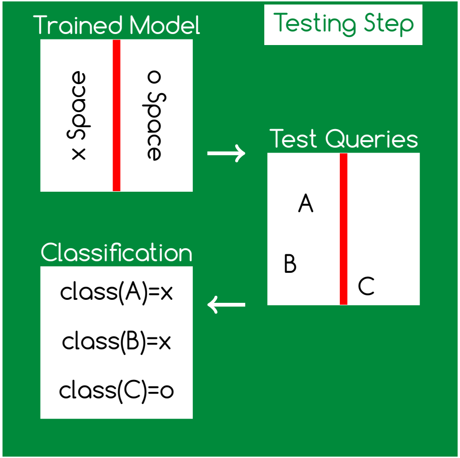
forbes.com

Outline

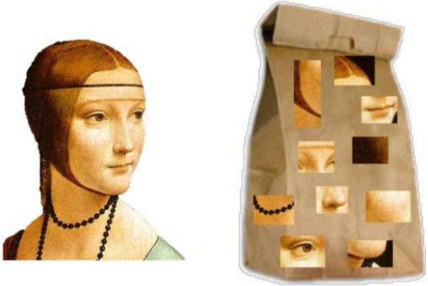
What is Machine Learning?

- ML is a field of algorithm development wherein data is used to tune parameters/weights towards some task (like classification). We are going to discuss **supervised** ML, meaning the data is labeled.



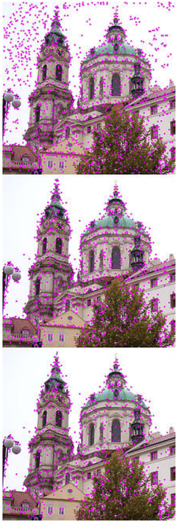


- Traditionally, **features** are extracted from data samples to focus the training/testing of the machine learning model.



MIT

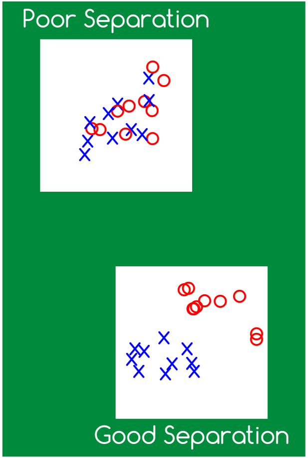
- Traditionally, **features** are extracted from data samples to focus the training/testing of the machine learning model.
- How these features are designed and the attributes they capture is of great interest; the more discriminatory they are, the easier a classifier will train and the better the algorithm will do.



Wikipedia

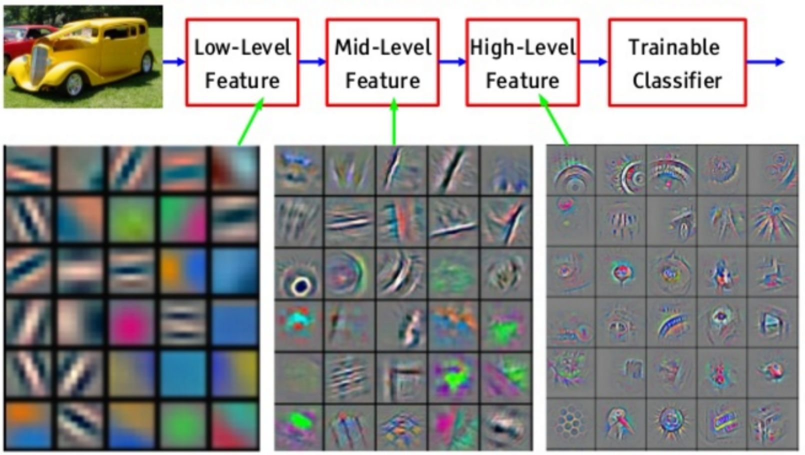
Features

- Traditionally, **features** are extracted from data samples to focus the training/testing of the machine learning model.
- How these features are designed and the attributes they capture is of great interest; the more discriminatory they are, the easier a classifier will train and the better the algorithm will do.



Deep Learning: Let the Machine Decide Features

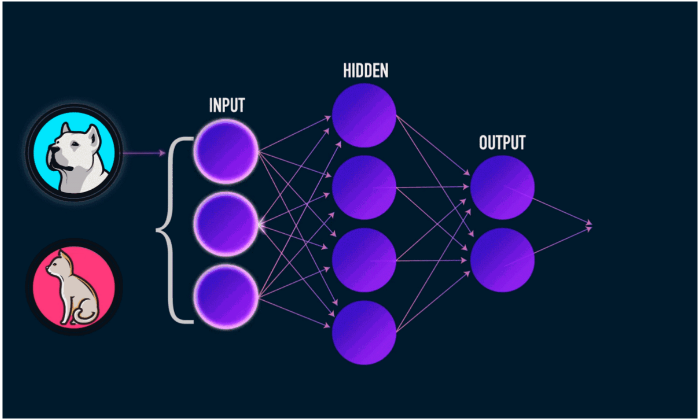
- Instead of human developed features, why not let the **machine decipher its own set of discriminatory features**.



Yan LeCunn

Deep Learning: Let the Machine Decide Features

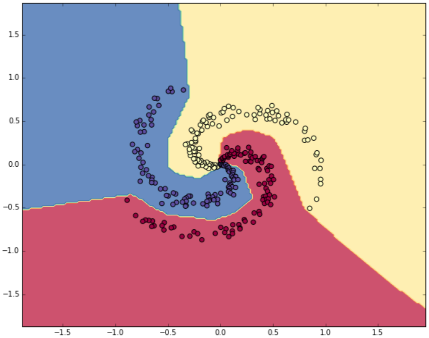
- Instead of human developed features, why not let the **machine decipher its own set of discriminatory features**.



towardsdatascience.com

Deep Learning: Let the Machine Decide Features

- Instead of human developed features, why not let the **machine decipher its own set of discriminatory features**.



Stanford CS 231

- The strength of deep learning/neural networks is in the auto-feature-generation. Circumventing human bias allows a direct path to a seemingly optimal solution.

Delving Deep into Rectifiers:
Surpassing Human-Level Performance on ImageNet Classification

Kaiming He Xiangyu Zhang Shaoqing Ren Jian Sun

Microsoft Research
{kahe, v-xiangz, v-shren, jiansun}@microsoft.com

Abstract

*Rectified activation units (rectifiers) are essential for state-of-the-art neural networks. In this work, we study rectifier neural networks for image classification from two aspects. First, we propose a Parametric Rectified Linear Unit (PReLU) that generalizes the traditional rectified unit. PReLU improves model fitting with nearly zero extra computational cost and little overfitting risk. Second, we derive a robust initialization method that particularly considers the rectifier nonlinearities. This method enables us to train extremely deep rectified models directly from scratch and to investigate deeper or wider network architectures. Based on our PReLU networks (PReLU-nets), we achieve **4.94%** top-5 test error on the ImageNet 2012 classification dataset. This is a 26% relative improvement over the ILSVRC 2014 winner (GoogLeNet, 6.66% [29]). To our knowledge, our result is the first to surpass human-level performance (5.1%, [22]) on this visual recognition challenge.*

and the use of smaller strides [33, 24, 2, 25]), new non-linear activations [21, 20, 34, 19, 27, 9], and sophisticated layer designs [29, 11]. On the other hand, better generalization is achieved by effective regularization techniques [12, 26, 9, 31], aggressive data augmentation [16, 13, 25, 29], and large-scale data [4, 22].

Among these advances, the rectifier neuron [21, 8, 20, 34], *e.g.*, Rectified Linear Unit (ReLU), is one of several keys to the recent success of deep networks [16]. It expedites convergence of the training procedure [16] and leads to better solutions [21, 8, 20, 34] than conventional sigmoid-like units. Despite the prevalence of rectifier networks, recent improvements of models [33, 24, 11, 25, 29] and theoretical guidelines for training them [7, 23] have rarely focused on the properties of the rectifiers.

In this paper, we investigate neural networks from two aspects particularly driven by the rectifiers. First, we propose a new generalization of ReLU, which we call

He *et al* CVPR 2015

Deep Learning & Its Features

- The strength of deep learning/neural networks is in the auto-feature-generation. Circumventing human bias allows a direct path to a seemingly optimal solution.
- Without human intervention, though, we have a **black box** classifier.

Query A → Network → class(A)

What is the network looking at?

Why those features?

Are the features relevant?

The Rest of this Talk...

- For the rest of this talk, we are going to discuss ways in which deep learning/neural networks can fooled.
- Key context: deep learning is **the state-of-the-art**. It is difficult to justify using SVMs/random forests/etc. for many problems when a neural network can *significantly* improve performance.
- Note as well that these other machine learning strategies can also be fooled - many in the exact same way.

Outline

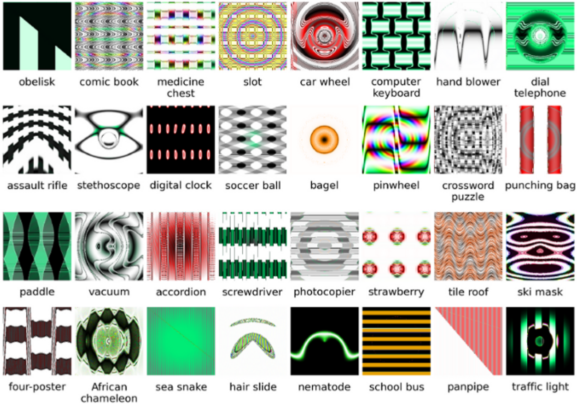
Setting

- Assume for the following that we have a well-trained neural network for a classification task.
- It is easiest to illustrate the following ideas with images, but note that they apply for any domain.

Outline

What Is the Network Looking For?

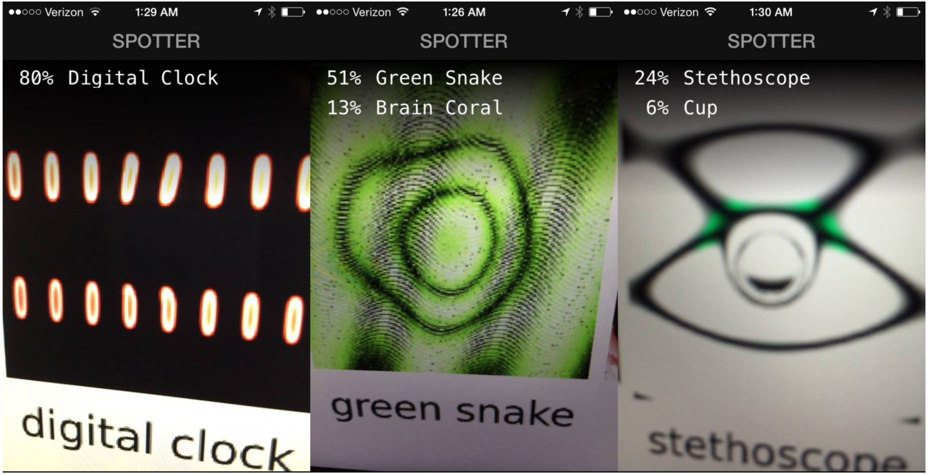
- We did not constrain the network to filters that we can understand.
- Research has shown that deformations of objects into gibberish can still earn high scores from a neural network. The images below all have $> 99\%$ confidence from a well-trained model.



Nguyen *et al* CVPR 2015

What Is the Network Looking For?

- Even second-hand, these images fool state-of-the-art networks.

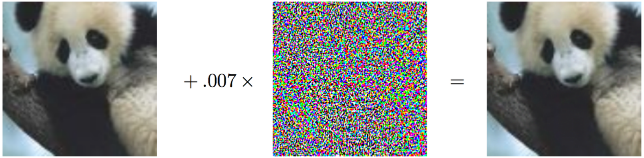


Nguyen *et al* CVPR 2015

Outline

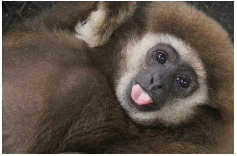
How robust is the model to being fooled?

- What if someone wants to actively fool a network? What if we have an **adversarial attack**?



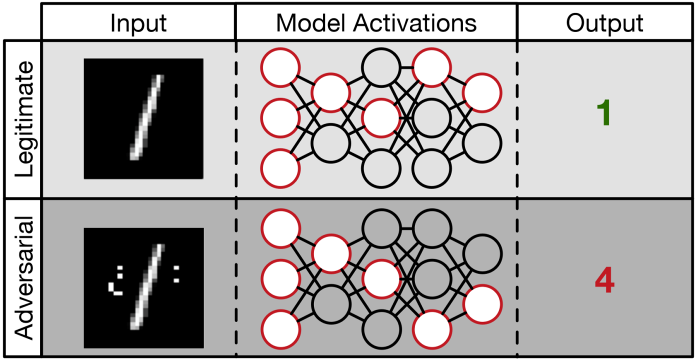
$\boldsymbol{x}$	$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$	$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"panda"	"nematode"	"gibbon"
57.7% confidence	8.2% confidence	99.3 % confidence

Figure 1: A demonstration of fast adversarial example generation applied to GoogLeNet (Szegedy et al., 2014a) on ImageNet. By adding an imperceptibly small vector whose elements are equal to the sign of the elements of the gradient of the cost function with respect to the input, we can change GoogLeNet's classification of the image. Here our ϵ of .007 corresponds to the magnitude of the smallest bit of an 8 bit image encoding after GoogLeNet's conversion to real numbers.

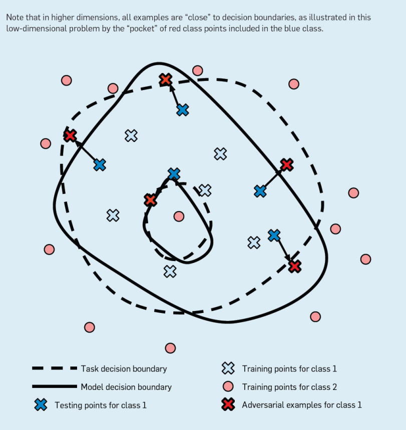


Goodfellow *et al* ICLR 2014

How Attacks Work



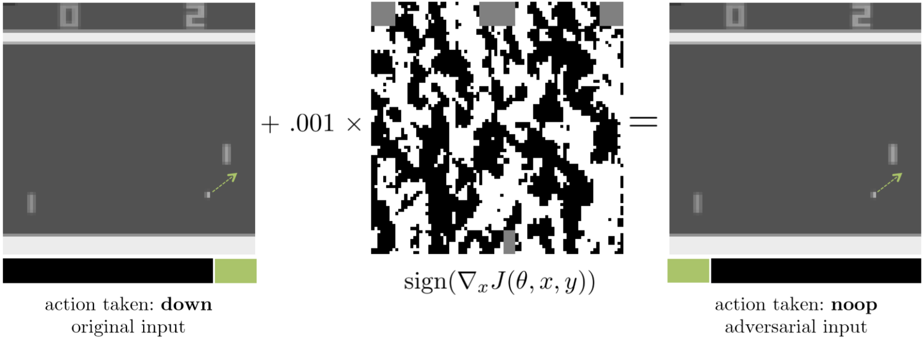
Papernot *et al* 2016 arXiv



GoodFellow *et al* ACM Magazine 2018

White Box Attacks

- White box attack:**the adversary knows the model parameters**.



Huang *et al* ICLR 2017

Black Box Attacks

- Black box attacks: **the adversary doesn't know model parameters**.
- These attacks are *harder* to deal with than white box attacks.

Model A, Model B, Model C $\rightarrow$ Utilize Model Parameters to Get Perturbations $\rightarrow$
Adversarial Attack Function $Q_A(\cdot)$
Adversarial Attack Function $Q_B(\cdot)$
Adversarial Attack Function $Q_C(\cdot)$

White Box Attack

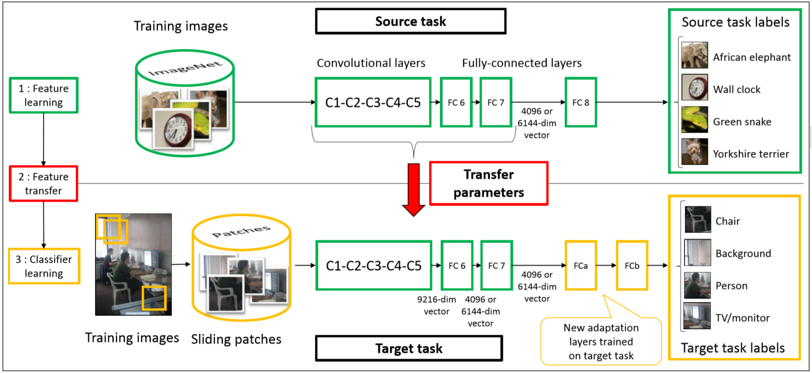
Test Query $X \longrightarrow$ Generate Adversary $Q_A(X) \longrightarrow \text{class}_A(Q_A(X)) \neq \text{class}_A(X)$

Black Box Attack

Query $X \rightarrow$ Generate Adversary $Q_A(X) \rightarrow$
$\text{class}_B(Q_A(X)) \neq \text{class}_B(X)$
$\text{class}_C(Q_A(X)) \neq \text{class}_C(X)$

Why Do Black Box Attacks Work?

- A key concept of modern NN theory is **transfer learning**, the ability to share weights among similar tasks.
- Sharing weights makes training with smaller data sets possible, but it also means that similar models will produce similar weights $\rightarrow$ black box attacks.



Oquab *et al* CVPR 2014

Outline

- We can prevent white box attacks by training with adversarial examples.
- Though simple, this is an effective measure.

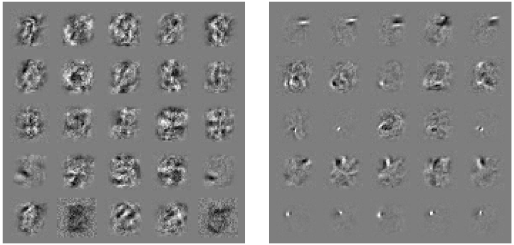
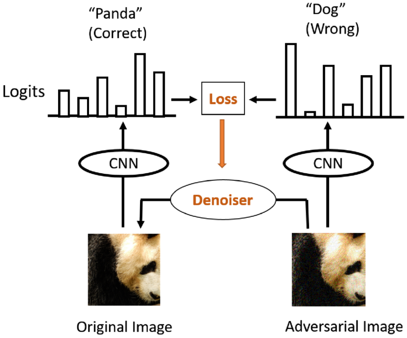


Figure 3: Weight visualizations of maxout networks trained on MNIST. Each row shows the filters for a single maxout unit. Left) Naively trained model. Right) Model with adversarial training.

Goodfellow *et al* ICLR 2015

Protection from Black Box Attacks

- Similar to white box, we can use train several models and train using a mixed batch of adversarial images.
- This seems to work but is unsatisfying; there are other more sophisticated actions to take, but this is an open question.



Liao *et al* CVPR 2018

Outline

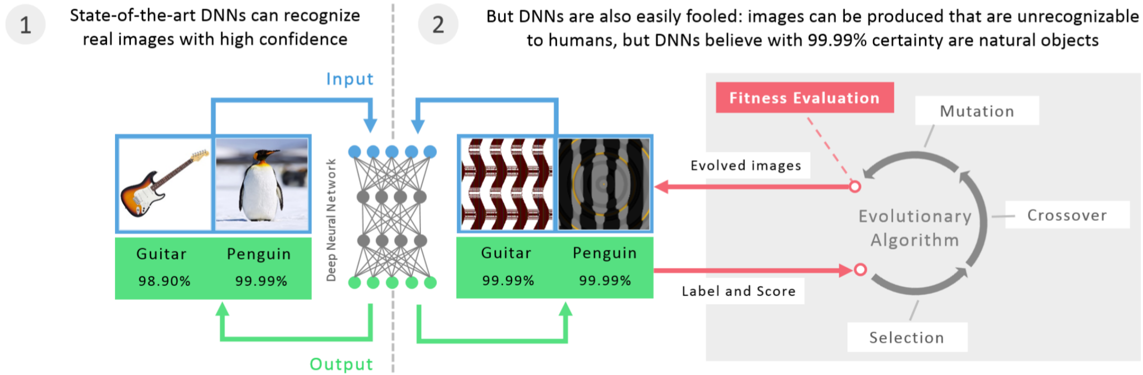
Summary

- Deep learning is the state-of-the-art. It's unavoidably the best choice for most classification tasks.
- It's a black box by design. We want the machine to craft its own features even though we won't be able to decipher their meaning.
- Adversarial images show the double edged sword of this feature generation. The incredible performance comes with vulnerabilities.
- For cybersecurity, we need to look into neural networks as an option but should be wary of their problem cases.

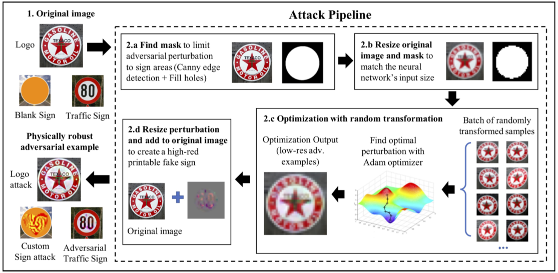
Outline

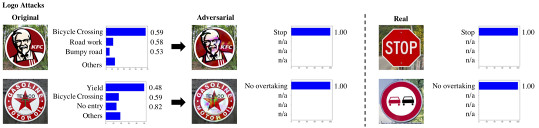
5 Appendix

How the Gibberish Images are Made

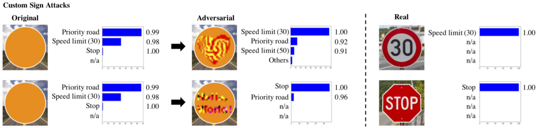


Nguyen *et al* CVPR 2015





(a) Classification of Logo attack examples. The adversarial examples are classified with high confidence as a real traffic sign, in spite of being out of the dataset.



(b) Classification of Custom Sign attack examples. The adversarial examples are classified with high confidence as a real traffic sign, in spite of being custom made signs.

Sitawarin *et al* ACM CCS 2018