ECE595 / STAT598: Machine Learning I

Lecture 33.1: Adversarial Attack - An Overview

Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

PURDUE
UNIVERSITY

Today's Agenda

- We have studied
  - Part 1: Basic learning pipeline
  - Part 2: Algorithms
  - Part 3: Learning theory
- Now, we want to study the robustness of learning algorithms
- Robustness = easiness to fail when input is perturbed. Perturbation can be in any kind.
- Robust machine learning is a very rich topic.
- In the past, we have robust SVM, robust kernel regression, robust PCA, etc.
- More recently, we have **transfer learning** etc.
- In this course, we will look at something very narrow, called **adversarial robustness**.
- That is, robustness against **attacks**.
- Adversarial attack is a very **hot** topic, as of today.
- We should not over-emphasize its importance. There are many other important problems.

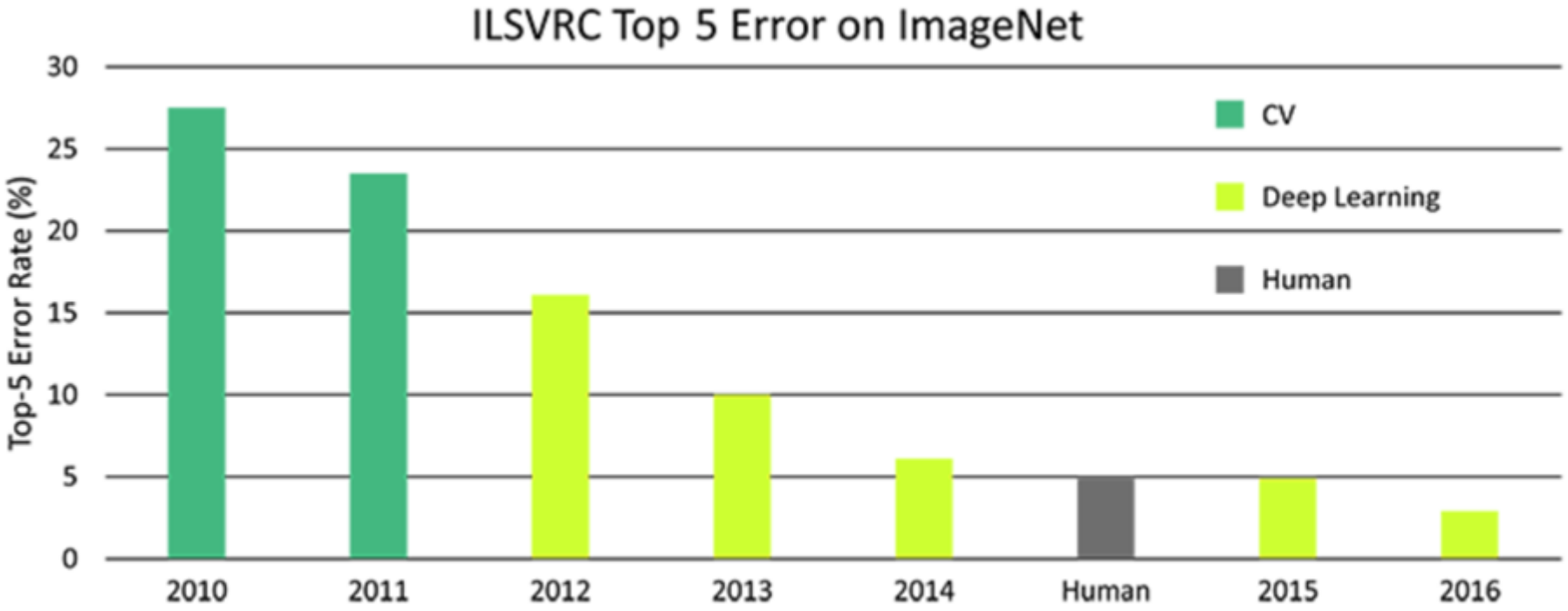
Outline

- Lecture 33 Overview
- Lecture 34 Min-distance attack
- Lecture 35 Max-loss attack and regularized attack

Today's Lecture

- What are adversarial attacks?
  - The surprising findings by Szegedy (2013) and Goodfellow (2014)
  - Examples of attacks
  - Physical attacks
- Basic terminologies
  - Defining attack
  - Multi-class problem
  - Three forms of attack
  - Objective function and constraint sets

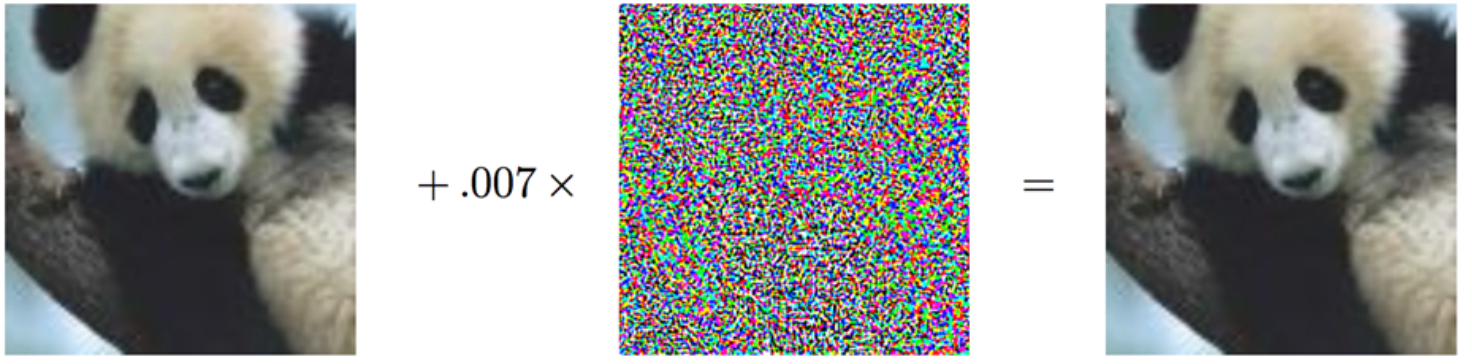
A Report in 2017



source: https://www.dsiac.org/resources/journals/dsiac/winter-2017-volume-4-number-1/real-time-situ-intelligent-video-analytics

Adversarial Attack Example: FGSM

- It is not difficult to fool a classifier
- The perturbation could be perceptually not noticeable



$\boldsymbol{x}$

"panda"
57.7% confidence

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

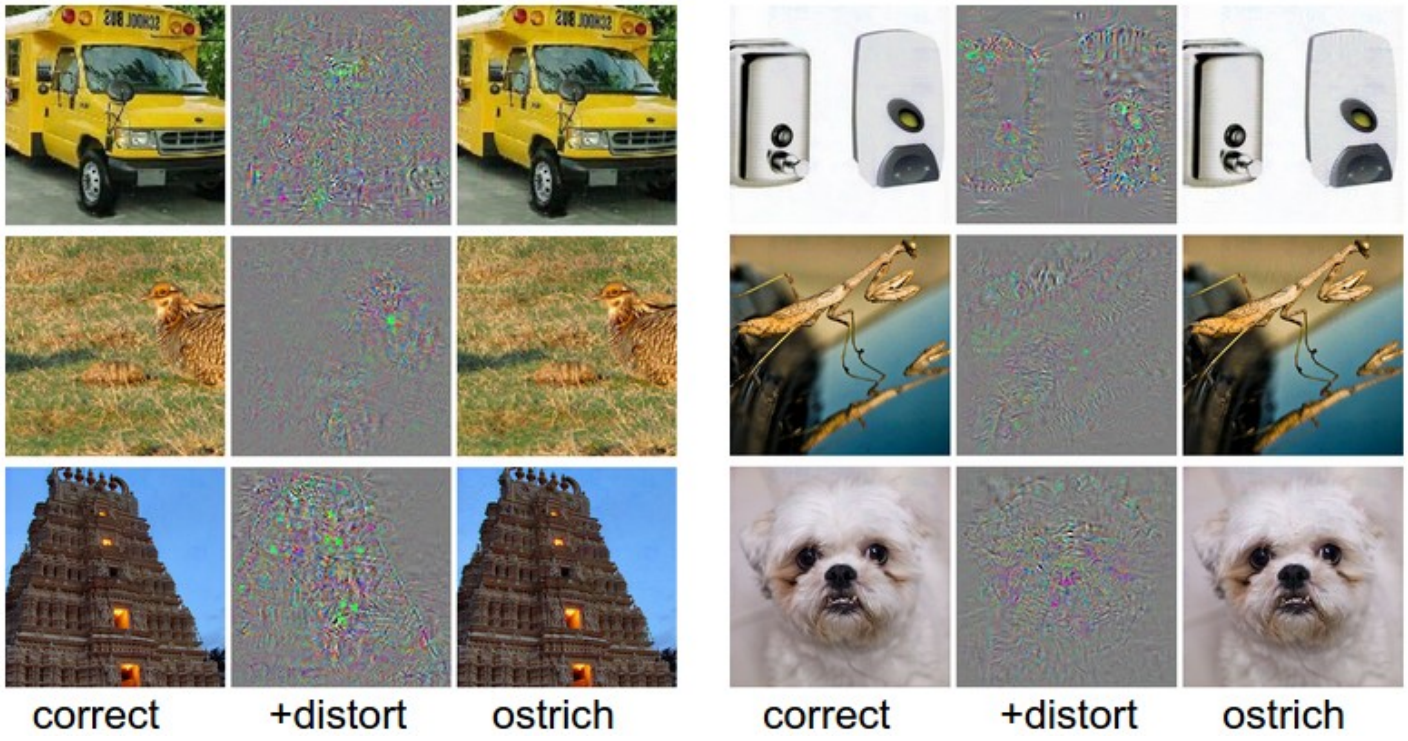
$\boldsymbol{x} + \epsilon \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
99.3 % confidence

Goodfellow et al. "Explaining and Harnessing Adversarial Examples", https://arxiv.org/pdf/1412.6572.pdf

Adversarial Attack Example: Szegedy's 2013 Paper

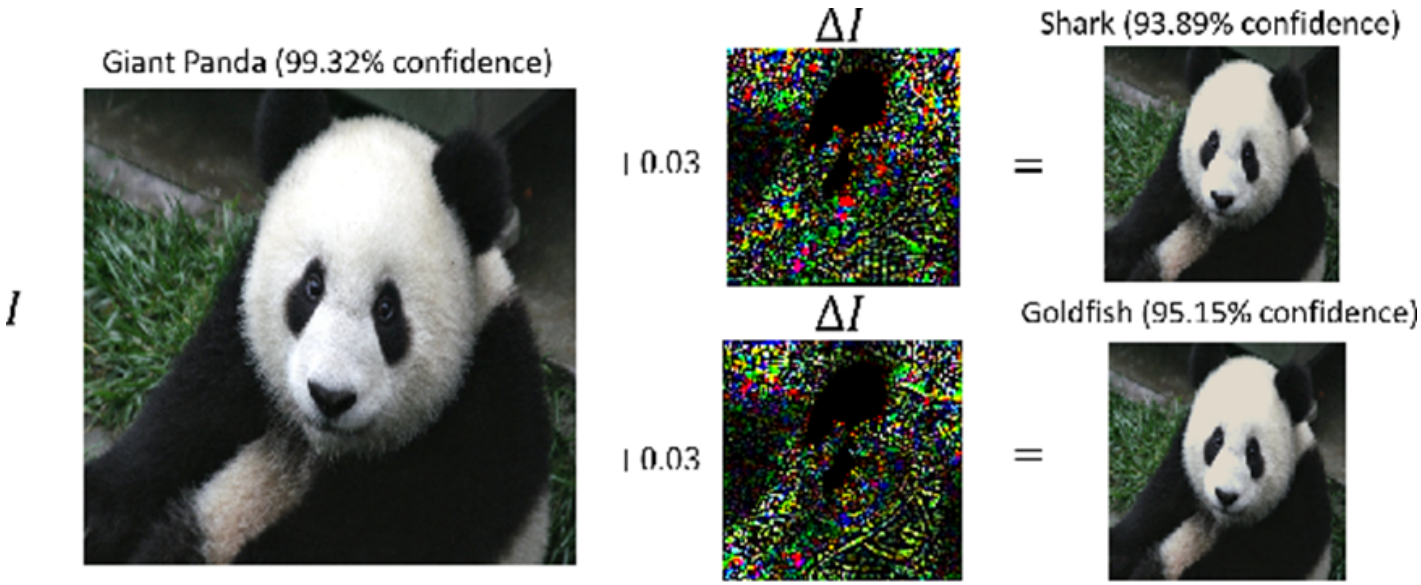
- This paper actually appears one year before Goodfellow's 2014 paper.



Szegedy et al. Intriguing properties of neural networks
https://arxiv.org/abs/1312.6199

Adversarial Attack: Targeted Attack

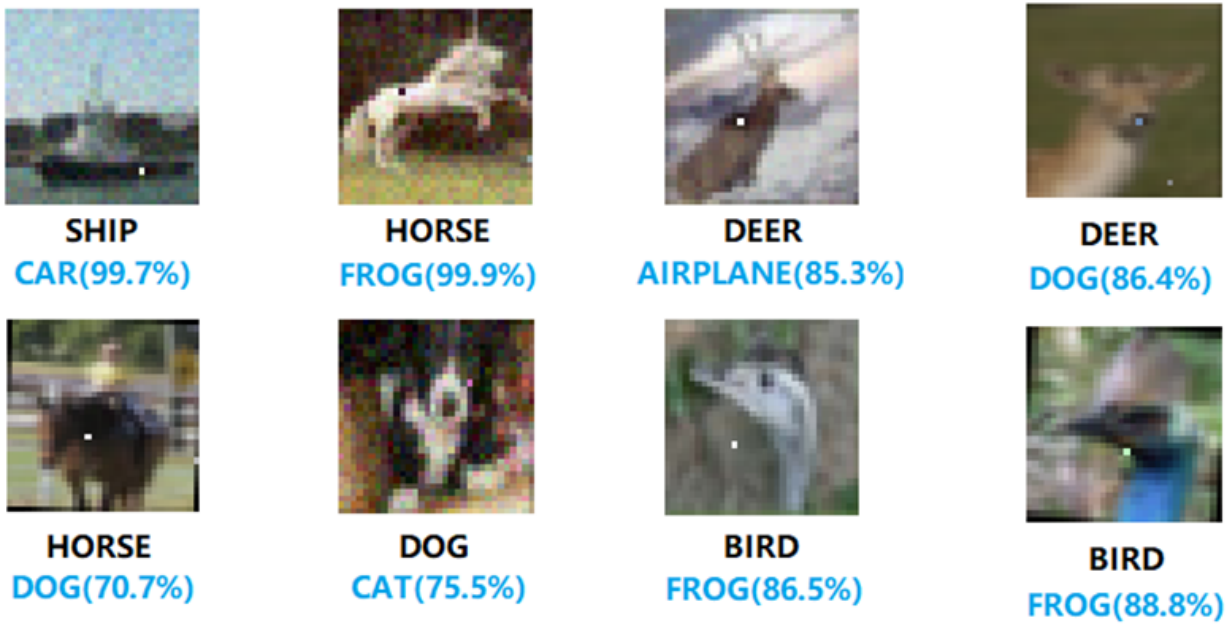
- Targeted Attack



Adversarial Examples Detection in Deep Networks with Convolutional Filter Statistics, https://arxiv.org/abs/1612.07767

Adversarial Attack Example: One Pixel

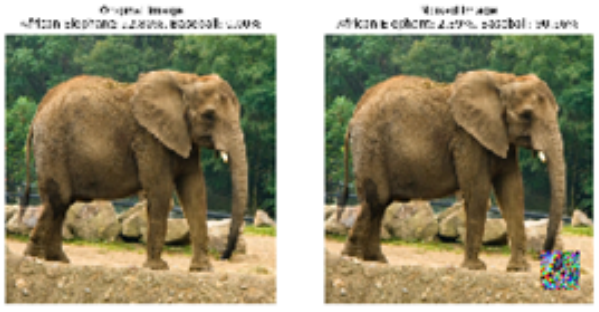
- One-pixel Attack



One pixel attack for fooling deep neural networks https://arxiv.org/abs/1710.08864

Adversarial Attack Example: Patch

- Adding a patch



African-Elephant (92.8%) → Baseball (90.7%)

Sports Car (92.8%) → Shih-Tzu (90.7%)

Brown Bear (87.9%) → **Tree Frog** (82.7%)

Minivan (90.7%) → **Tree Frog** (86.4%)

LaVAN: Localized and Visible Adversarial Noise, https://arxiv.org/abs/1801.02608

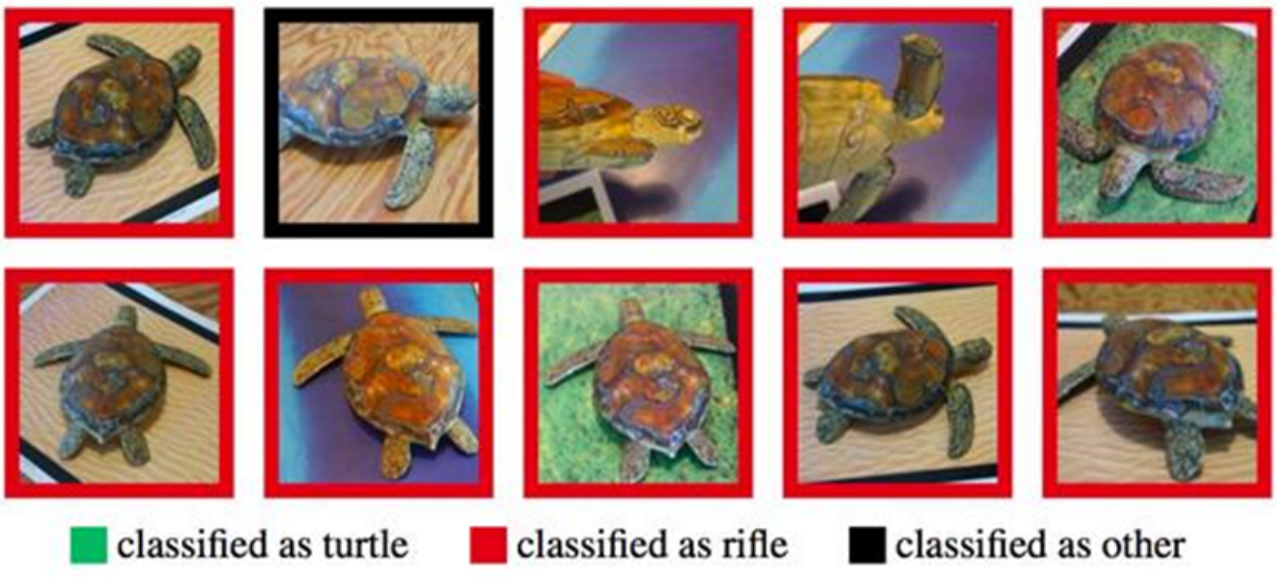
Adversarial Attack Example: Stop Sign

- The Michigan / Berkeley Stop Sign

Robust Physical-World Attacks on Deep Learning Models
https://arxiv.org/abs/1707.08945

Adversarial Attack Example: Turtle

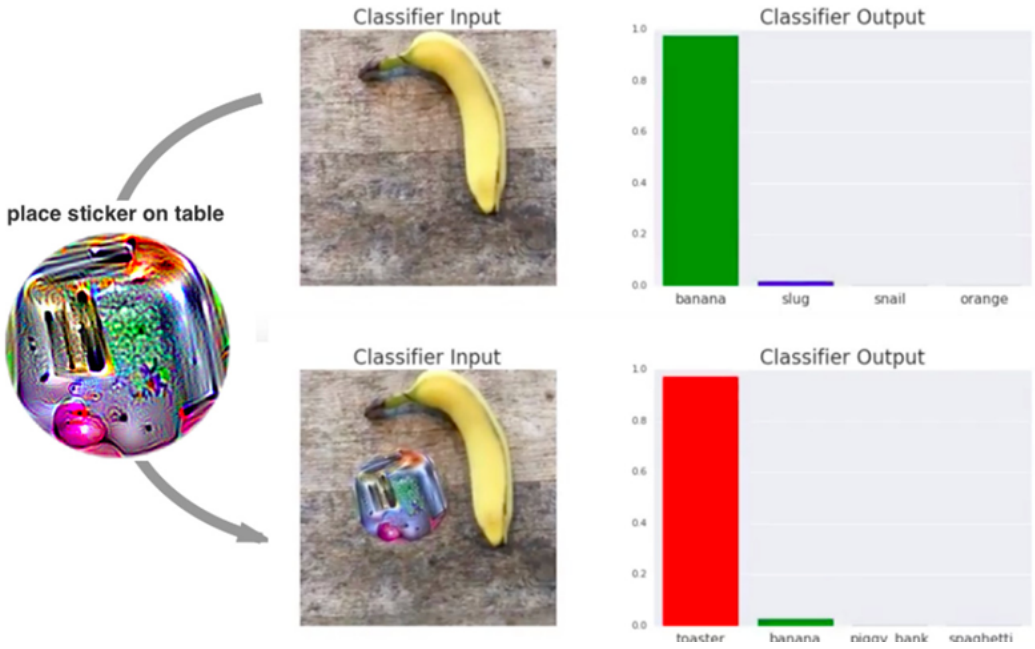
- The MIT 3D Turtle



Synthesizing Robust Adversarial Examples
https://arxiv.org/pdf/1707.07397.pdf
https://www.youtube.com/watch?v=YXy6oX1iNoA

Adversarial Attack Example: Toaster

- Google Toaster



Adversarial Patch
https://arxiv.org/abs/1712.09665
https://www.youtube.com/watch?v=i1sp4X57TL4

Adversarial Attack Example: Glass

- CMU Glass

Sharif, M., Bhagavatula, S., Bauer, L., & Reiter, M. K. (2016, October).
Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition.
In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (pp. 1528-1540). ACM.

Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition
`https://www.cs.cmu.edu/~sbhagava/papers/face-rec-ccs16.pdf`
`https://www.archive.ece.cmu.edu/~lbauer/proj/advml.php`

Adversarial Attack: A Survey in 2017

Table III: Summary of Applications for Adversarial Examples

Applications	Representative Study	Method	Adversarial Falsification	Adversary's Knowledge	Adversarial Specificity	Perturbation Scope	Perturbation Limitation	Attack Frequency	Perturbation Measurement	Dataset	Architecture
Reinforcement Learning	[93]	FGSM	N/A	White-box & Black-box	Non-Targeted	Individual	N/A	One-time	$\ell_1, \ell_2, \ell_\infty$	Atari	DQN, TRPO, A3C
	[94]	FGSM	N/A	White-box	Non-Targeted	Individual	N/A	One-time	N/A	Atari Pong	A3C
Generative Modeling	[95]	Feature Adversary, C&W	N/A	White-box	Targeted	Individual	Optimized	Iterative	ℓ_2	MNIST, SVHN, CelebA	VAE, VAE-GAN
	[96]	Feature Adversary	N/A	White-box	Targeted	Individual	Optimized	Iterative	ℓ_2	MNIST, SVHN	VAE, AE
Face Recognition	[67]	Impersonation & Dodging Attack	False negative	white-box & black-box	Targeted & Non-Targeted	Universal	Optimized	Iterative	Total Variation	LFW,	VGGFace
Object Detection	[22]	DAG	False negative & False positive	White-box & Black-box	Non-Targeted	Individual	N/A	Iterative	N/A	VOC2007, VOC2012	Faster-RCNN
Semantic Segmentation	[22]	DAG	False negative & False positive	White-box & Black-box	Non-Targeted	Individual	N/A	Iterative	N/A	DeepLab	FCN
	[97]	ILLC	False negative	White-box	Targeted	Individual	N/A	Iterative	ℓ_∞	Cityscapes	FCN
	[98]	ILLC	False negative	White-box	Targeted	Universal	N/A	Iterative	N/A	Cityscapes	FCN
Reading Comprehension	[99]	AddSent, AddAny	N/A	Black-box	Non-Targeted	Individual	N/A	One-time & Iterative	N/A	SQuAD	BiDAF, Match-LSTM, and twelve other published models
	[100]	Reinforcement Learning	False negative	White-box	Non-Targeted	Individual	Optimized	Iterative	ℓ_0	TripAdvisor Dataset	Bi-LSTM, memory network
Malware Detection	[101]	JSMA	False negative	White-box	Targeted	Individual	Optimized	Iterative	ℓ_2	DREBIN	2-layer FC
	[102]	Reinforcement Learning	False negative	Black-box	Targeted	Individual	N/A	Iterative	N/A	N/A	Gradient Boosted Decision Tree
	[103]	GAN	False negative	Black-box	Targeted	Individual	N/A	Iterative	N/A	malwr	Multi-layer Perceptron
	[104]	GAN	False negative	Black-box	Targeted	Individual	N/A	Iterative	N/A	Alexa Top 1M	Random Forest
	[105]	Generic Programming	False negative	Black-box	Targeted	Individual	N/A	Iterative	N/A	Contagio	Random Forest, SVM

Adversarial Examples: Attacks and Defenses for Deep Learning
https://arxiv.org/abs/1712.07107