Deep Learning: Vulnerabilities, Defenses and Beyond

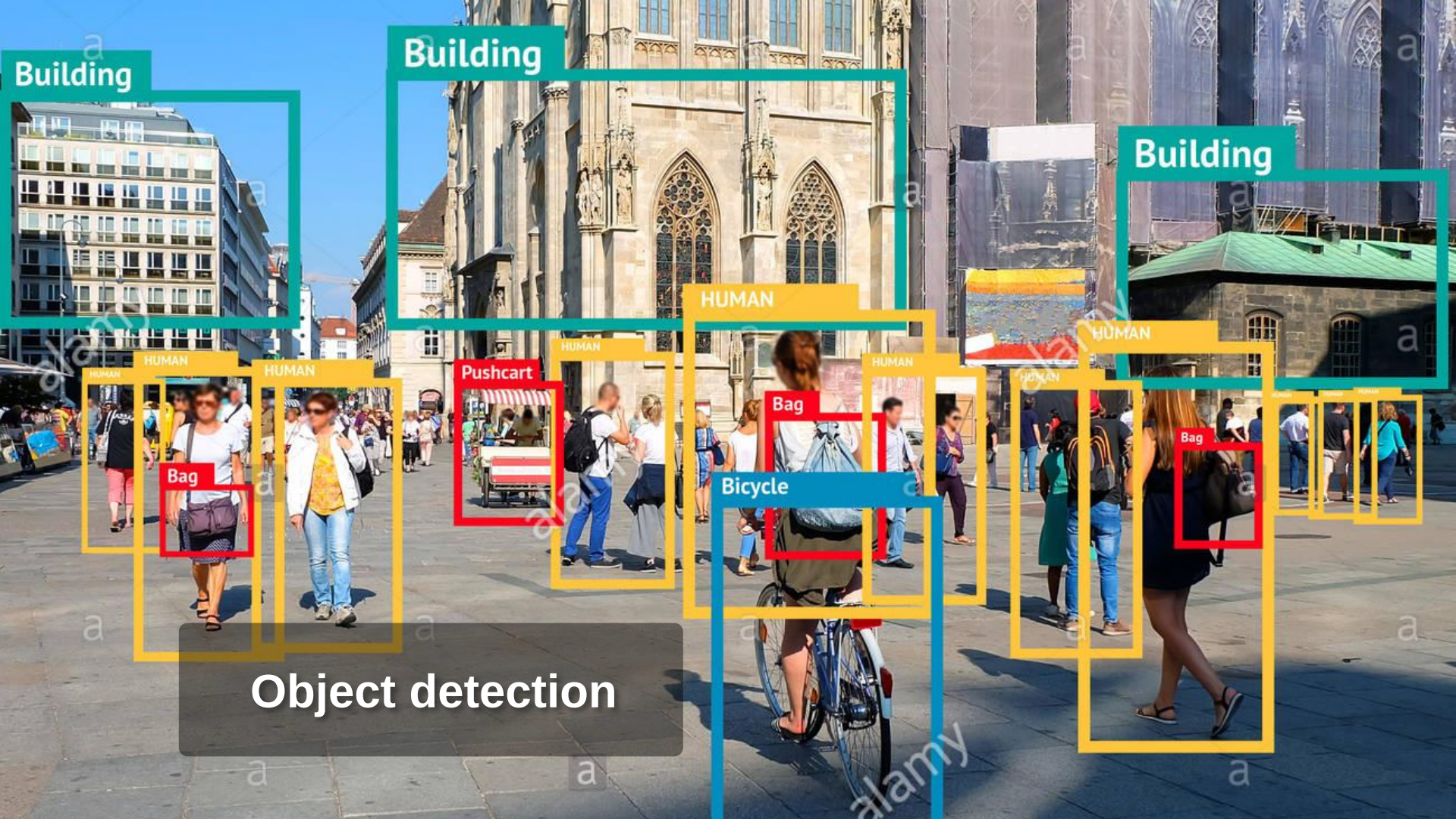
Prof. Yossi Keshet

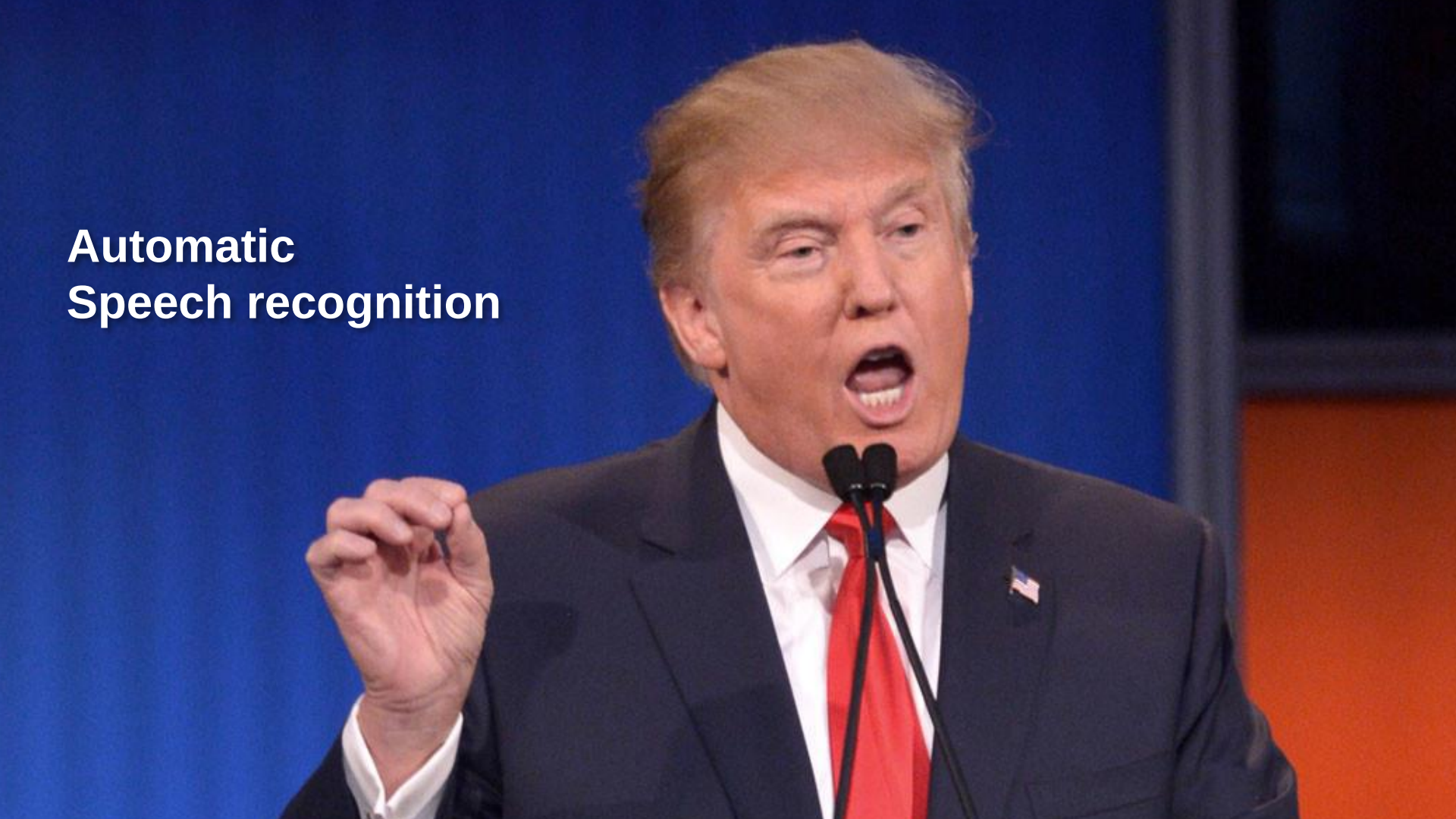
CS@BIU DEPARTMENT OF COMPUTER SCIENCE BAR-ILAN UNIVERSITY

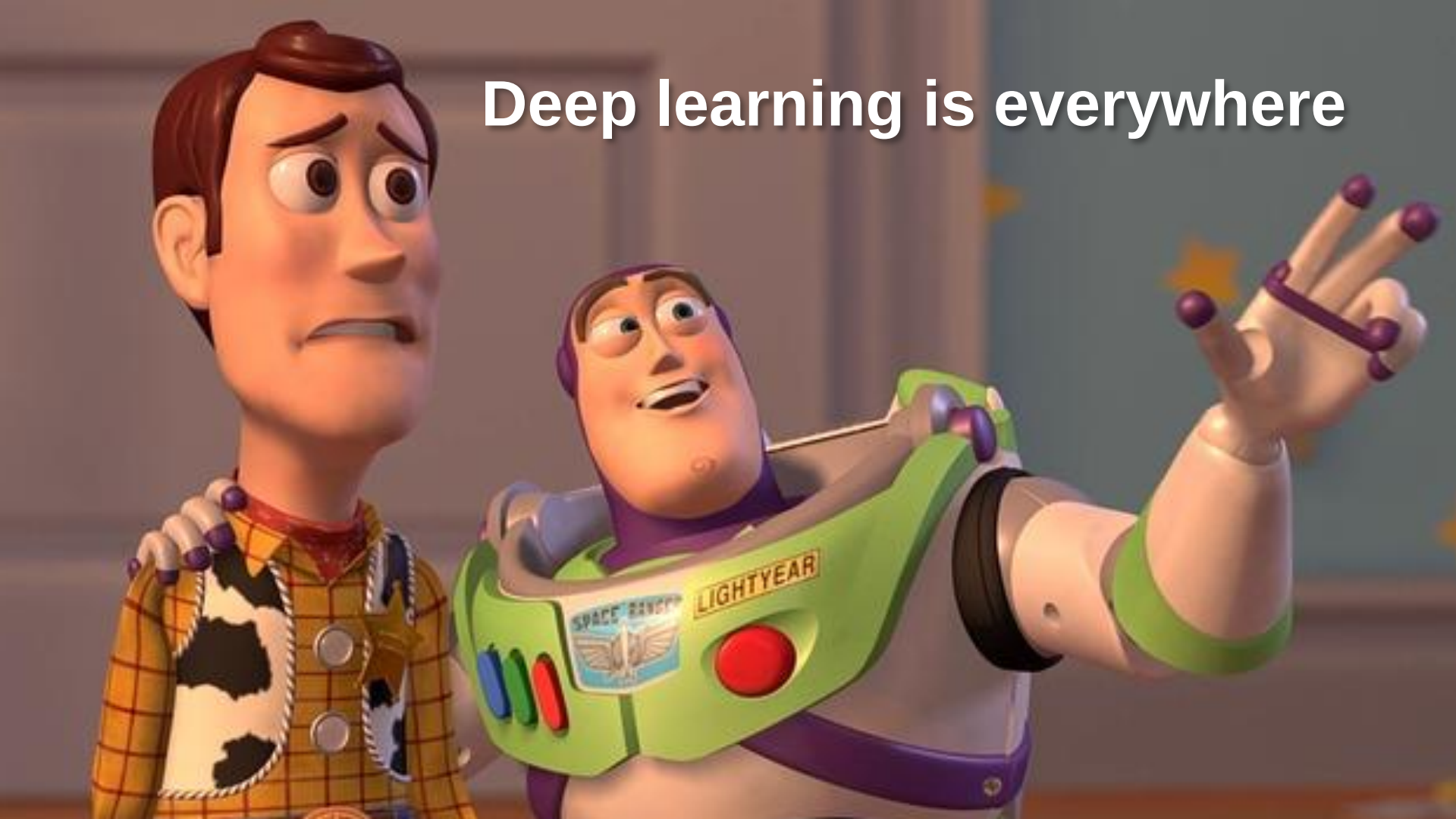
Since 2013 deep neural networks
have match human performance at…


Face detection


Street address detection


Building
Building
Building
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
Bag
Bag
Bag
Pushcart
Bicycle
Object detection


Automatic
Speech recognition


Deep learning is everywhere
LIGHTYEAR

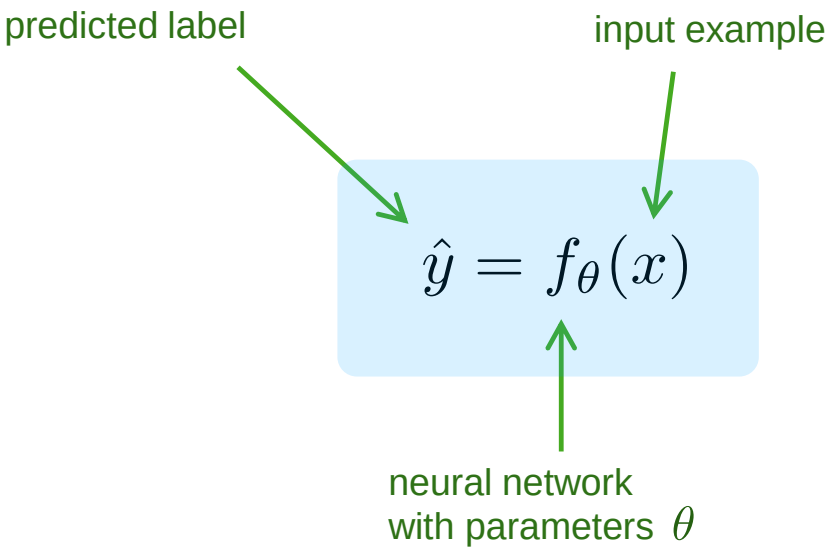

… but how secured are deep learning models?

Outline

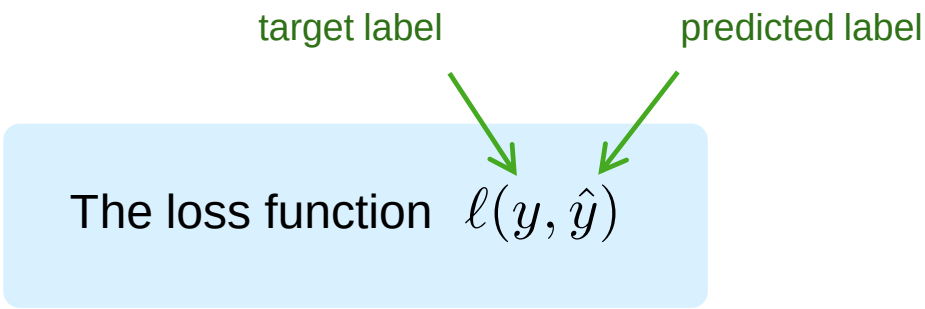
- **Deep learning review**
- **Adversarial examples**
  - Introduction
  - Houdini: generalize to any machine learning algorithm
    Cisse, Adi, Neverova, & Keshet (2017)
  - Speaker verification attack
    Kreuk, Adi, Cisse, & Keshet, (2017)
  - Adversarial Malware
    Kreuk, Barak, Aviv-Reuven, Baruch, Pinkas & Keshet (2018)
- **Watermarking machine learning models**
  Adi, Baum, Cisse, Pinkas, and Keshet (2018)
- **Defenses and detection of adversarial attacks**
  Shalev, Adi, and Keshet (2018)

Deep learning review

Deep learning model



Measuring performance



Measuring performance

– “0-1” loss in binary classification $$\ell(y, \hat{y}) = \begin{cases} 1 & y \neq \hat{y} \\ 0 & y = \hat{y} \end{cases}$$

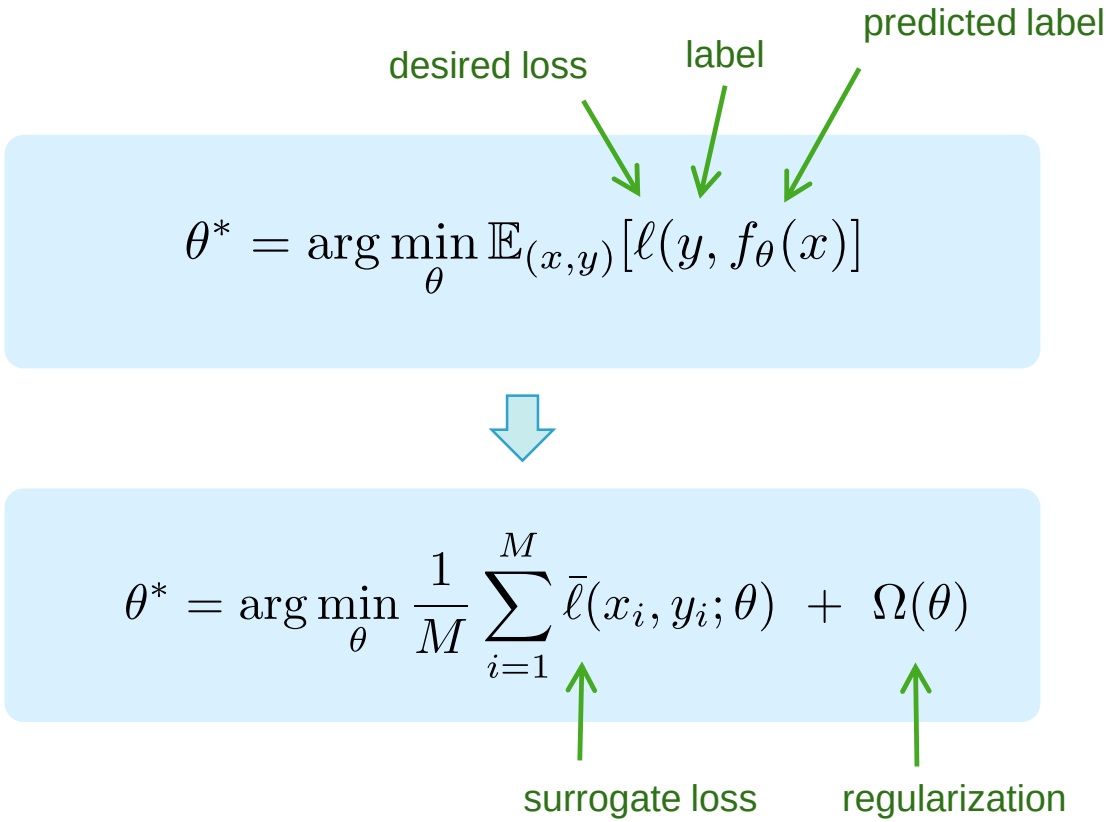
– Word Error Rate in speech recognition **It is easy to recognize speech**

It is easy to wreck a nice beach

1 substitutions
3 insertions

– Intersection-over-union in object segmentation

Training


desired loss
label
predicted label
$\theta^* = \arg\min_\theta \mathbb{E}_{(x,y)}[\ell(y, f_\theta(x)]$
$\theta^* = \arg\min_\theta \frac{1}{M} \sum_{i=1}^{M} \bar{\ell}(x_i, y_i; \theta) + \Omega(\theta)$
surrogate loss
regularization

Surrogate loss functions

- Negative log-likelihood:

$$\bar{\ell}_{NLL}(x, y, \theta) = -\log \mathbb{P}(x, y|\theta)$$

- Hinge loss function (similar to SVM):

$$\bar{\ell}_{hinge}(x, y, \theta) = \ell(y, \hat{y}) - \mathbb{P}(x, y|\theta) + \mathbb{P}(x, \hat{y}|\theta)$$

- And there are more...

Optimization for training

$$\theta^* = \arg\min_{\theta} \sum_{i=1}^{M} \bar{\ell}(x_i, y_i; \theta) \; + \; \Omega(\theta)$$

The optimum is found using (stochastic) gradient descent method:

for $t = 1 \ldots T$

  pick example (x_j, y_j) uniformly at random

  compute gradient $g = \nabla \bar{\ell}\Big(f_\theta(x_j), y_j\Big)$

  update parameters $\theta \leftarrow \theta + \eta\, g$

Adversarial example attacks

deep learning model

PANDA

Goodfellow, Shlens & Szegedy (2015)

deep learning model

PANDA

Goodfellow, Shlens & Szegedy (2015)

deep learning model

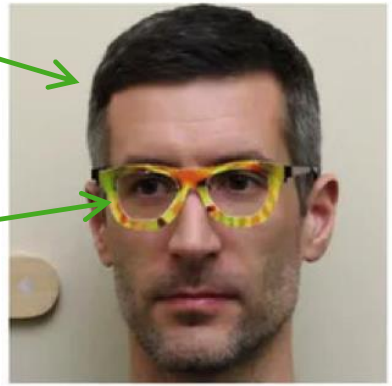
GIBBON

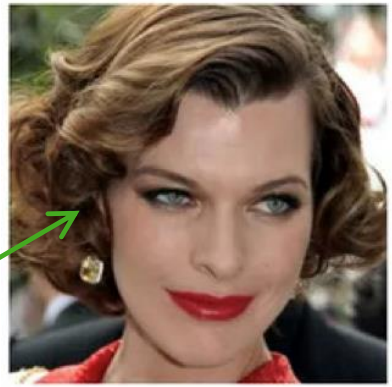
Goodfellow, Shlens & Szegedy (2015)

impersonator

adversarial perturbation

the real Milla Jovovich



Sharif, Bhagavatula, Bauer, & Reiter (2016)

deep learning model

STOP SIGN

Eykholt, et al. (2018)

deep learning model

Eykholt, et al. (2018)

Attacking Remotely Hosted Black-Box Models

Remote Platform	ML technique	Number of queries	Adversarial examples misclassified (after querying)
MetaMind	Deep Learning	6,400	84.24%
amazon web services™	Logistic Regression	800	96.19%
Google Cloud Platform	Unknown	2,000	97.72%

All remote classifiers were trained on MNIST data set (60,000 training examples and 10 classes)

Papernot, McDaniel, Goodfellow, Jha, Celik, & Swami (2017)

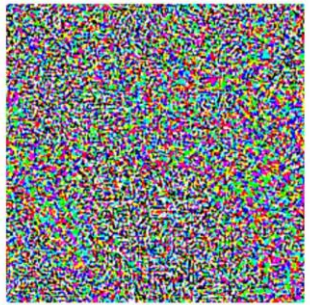
How the adversarial example is found?

$$\tilde{x} = x + \eta$$

adversarial example

input pattern

$+ .007 \times$



perturbation

How the perturbation is found?

$$\eta = \arg \max_{\eta:\|\eta\|_p \leq \epsilon} \bar{\ell}(x+\eta, y; \theta)$$

adversarial example $\tilde{x} = x + \eta$

Take first order Taylor Expansion:

$$\eta = \arg \max_{\eta:\|\eta\|_p \leq \epsilon} \left(\nabla_x \bar{\ell}(x, y; \theta)\right)^\top \eta$$

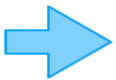
where $g = \nabla_x \bar{\ell}(x, y; \theta)$

How the perturbation is found?

Solving $\eta = \arg \max_{\eta:\|\eta\|_p \leq \epsilon} \left(\nabla_x \bar{\ell}(x, y; \theta) \right)^\top \eta$

where $g = \nabla_x \bar{\ell}(x, y; \theta)$

$\tilde{x} = x + \eta$



$\tilde{x} = x + \epsilon \cdot \text{sign}(g) \qquad p = \infty$

$\tilde{x} = x + \epsilon \cdot g \qquad p = 2$

Main result:

We showed that almost any machine learning-based model can be attacked, including models solving complex and structured tasks.

Cisse, Adi, Neverova, & Keshet (2017)

Deep networks for structured task

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} \; g_\theta(x, y)$$

the set of possible labels is exponentially large

Deep networks for structured task

Example: machine translation

Machine translation is awesome!

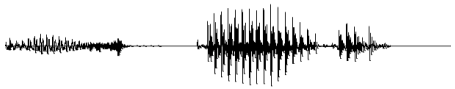
機械翻訳は素晴らしいです！

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} \; g_\theta(x, y)$$

the set of possible labels is exponentially large

Deep networks for structured task

Example: speech recognition



音声認識

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} \; g_\theta(x, y)$$

the set of possible labels is exponentially large

A new loss function called Houdini

$$\ell_H(x, y; \theta) = \mathbb{P}_{\gamma \sim \mathcal{N}(0,1)} \left[\gamma < g_\theta(x, \hat{y}) - g_\theta(x, y) \right] \cdot \ell(y, \hat{y})$$

Compare to the negative log-likelihood:

$$\bar{\ell}_{NLL}(x, y, \theta) = -\log \mathbb{P}(x, y|\theta)$$

Cisse, Adi, Neverova, & Keshet (2017)

Houdini properties I

Strong consistency: the model found using Houdini yields the infimum loss achievable by any predictor

$$\lim_{m\to\infty} \bar{\ell}_H(x, y; \theta_m) = \inf_{\theta} \ \mathbb{E}_{(x,y)\sim\rho}\Big[\ell(y, \hat{y}_\theta(x))\Big]$$

Cisse, Adi, Neverova, & Keshet (2017)

Houdini properties II

Lower bound to the loss

$$\ell_H(x, y; \theta) = \mathbb{P}_{\gamma \sim \mathcal{N}(0,1)} \left[\gamma < g_\theta(x, \hat{y}) - g_\theta(x, y)\right] \cdot \ell(y, \hat{y}) \leq \ell(y, \hat{y})$$

Cisse, Adi, Neverova, & Keshet (2017)

Houdini properties III

Gradients can be found analytically

$$\nabla_x \bar{\ell}_H(x, y; \theta) = \frac{\partial \bar{\ell}_H(x, y; \theta)}{\partial g_\theta(x, y)} \frac{\partial g_\theta(x, y)}{\partial x}$$

$$\nabla_g \left[\mathbb{P}_{\gamma \sim \mathcal{N}(0,1)} \left[\gamma < \overbrace{g_\theta(x, \hat{y}) - g_\theta(x, y)}^{\delta g(\hat{y}, y)} \right] \ell(y, \hat{y}) \right]$$

$$= \nabla_g \left[\frac{1}{\sqrt{2\pi}} \int_{\delta g(\hat{y}, y)}^{\infty} e^{-v^2/2} dv \right] \ell(y, \hat{y})$$

Cisse, Adi, Neverova, & Keshet (2017)

Houdini properties IV

The Houdini loss function

$$\ell_H(x, y; \theta) = \mathbb{P}_{\gamma \sim \mathcal{N}(0,1)} \left[\gamma < g_\theta(x, \hat{y}) - g_\theta(x, y)\right] \cdot \ell(y, \hat{y})$$

is very similar to the generalized Probit loss function:

$$\bar{\ell}_{\text{probit}}(x, y, \theta) = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,I)} \Big[\ell(y, \hat{y}_{\theta+\epsilon}(x))\Big]$$

Cisse, Adi, Neverova, & Keshet (2017)

Image segmentation: example 1

Cisse, Adi, Neverova, & Keshet (2017)

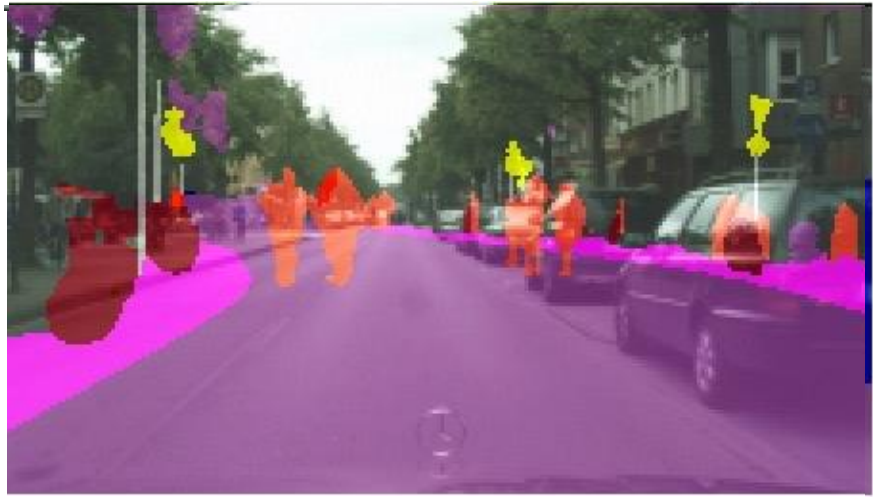
Image segmentation: example 1

Cisse, Adi, Neverova, & Keshet (2017)

Image segmentation: example 2

Cisse, Adi, Neverova, & Keshet (2017)

Image segmentation: example 2



Cisse, Adi, Neverova, & Keshet (2017)

$\tilde{x}$ η

(Cisse, Adi, Neverova, & Keshet, 2017)

Image segmentation: example 3

(Cisse, Adi, Neverova, & Keshet, 2017)

Image segmentation: example 3

(Cisse, Adi, Neverova, & Keshet, 2017)

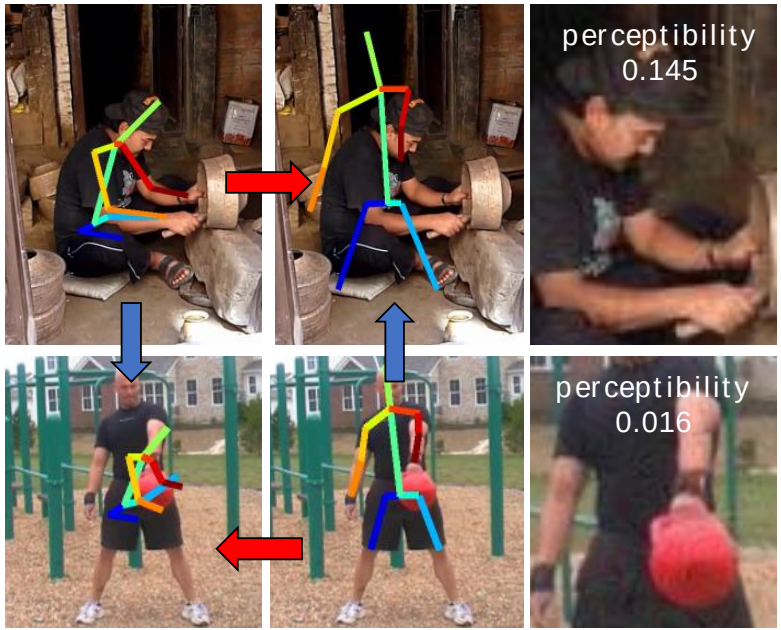
Image segmentation: example 3

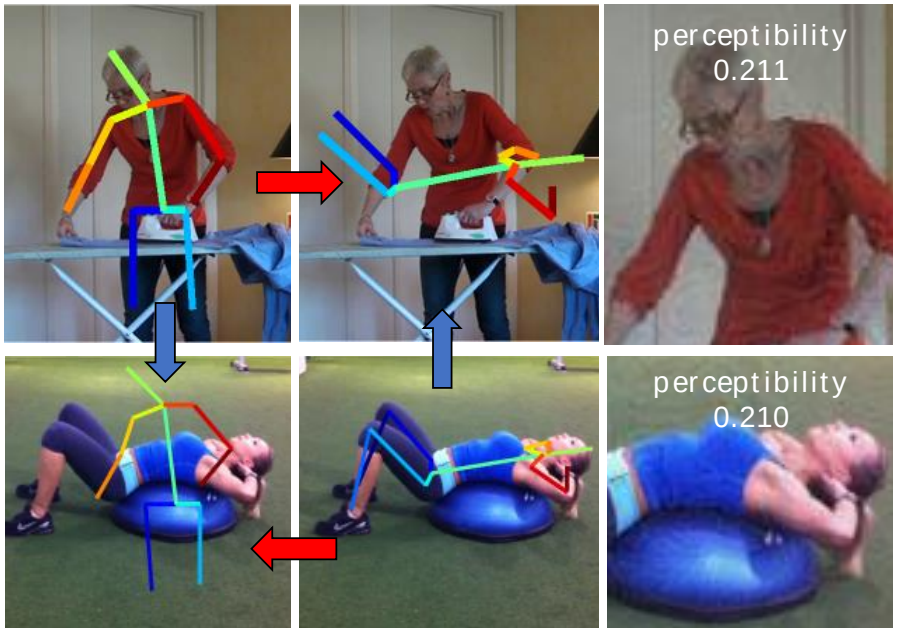
(Cisse, Adi, Neverova, & Keshet, 2017)

Image segmentation: example 3

(Cisse, Adi, Neverova, & Keshet, 2017)

Pose estimation (Xbox kinect): example 2

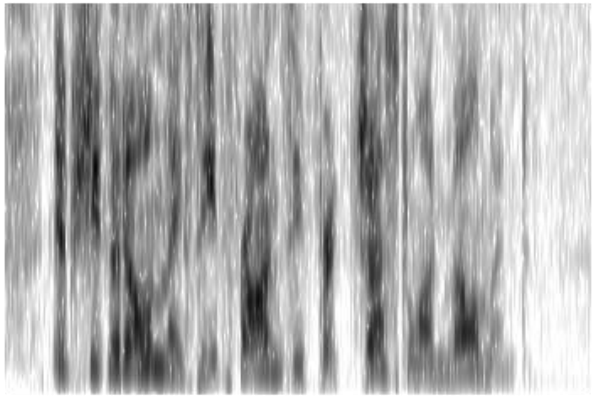




Cisse, Adi, Neverova, & Keshet (2017)

Speech recognition (Google Voice)

Original:
if she could only see Phronsie for just one moment



Adversarial:
if she ou down take shee throwns purhdress luon ellwon

Cisse, Adi, Neverova, & Keshet (2017)

60151999

MIT Technology Review

Intelligent Machines

AI Shouldn't Believe Everything It Hears

A new trick can fool voice-recognition systems into totall
mishearing what a recording says.

by Jamie Condliffe July 28, 2017

Artificial intelligence can accurately identify obje
image or recognize words uttered by a human,
algorithms don't work the same way as the hun
and that means that they can be spoofed in ways that hum

New Scientist reports that researchers from Bar-Ilan Univ
Israel and Facebook's AI team have shown that it's possibl
tweak audio clips so that a human understands them as n
voice-recognition AI hears something totally different. Th
works by adding a quiet layer of noise to a sound clip that
contains distinctive patterns a neural network will associa
words.

The team applied its new algorithm, called Houdini, to a s

New Scientist

HOME NEWS TECHNOLOGY SPACE PHYSICS HEALTH EARTH HUMANS LIFE TOPICS EVENTS JOBS

DAILY NEWS 27 July 2017

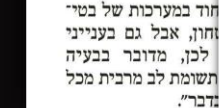
Sneaky attacks trick AIs into seeing or hearing what's not there

Open to hack attacks?
Asahi Shimbun via Getty Images

By Matt Reynolds

When it comes to AI, seeing isn't always believing. It's possible to trick machine learning systems into hearing and seeing things that aren't really there.

We already know that wearing a pair of spazzy glasses can fool face recognition software into thinking you're



גם על המכו
חכמה אי אפ

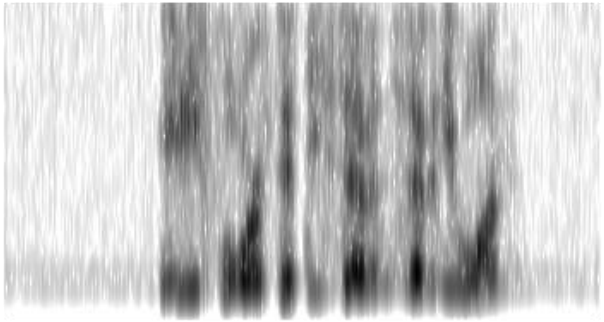
חוקרים ישראלים גילו דרך להערים על
גוגל ולהטעות קונסטולת משחקים של מ

עידו אפרתי

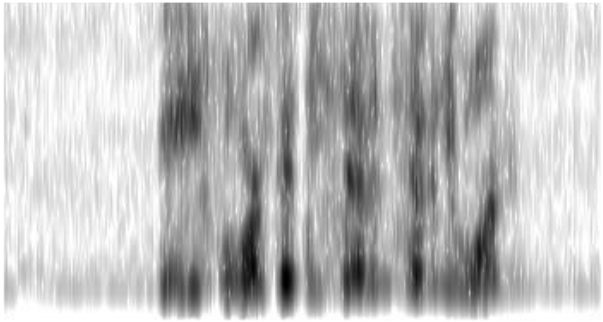
בסרטוני ההדגמה של המ
כונית האוטונומית מקפידים
הכמאים להתמקד ברגע אחה
מווקק וסמלי: השנייה שבה
האדם הנוסע לראשונה ברכב
ללא נהג שומט בחשש רב את
ידיו מן ההגה, תוך כדי נסיעה.
ומעביר את המושכות למחשב.
וזו אולי מפגן היכולות המ
רשים ביותר עד כה של בינה
מלאכותית, או למידת מכונה
(Machine Learning) במ
ציאות היום יומית. כל פעולה
או תגובה של המכונית דורשת
יכולות עיבוד חישוביות מתקד
מות. לעתים הצלחותיה משכנ
עות עד כדי כך, שניתן בטעות
לחשוב שהמכונית חווה ומנתחת
את המציאות כמעט כמו מוח
אנושי.

אבל מחקר שערכו חוקרים
ישראלים חשף באחרונה חול

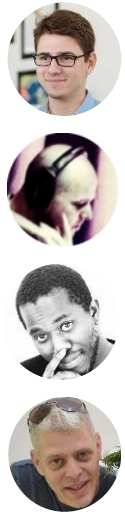
Speaker verification (YOHO)



Original: speaker 148

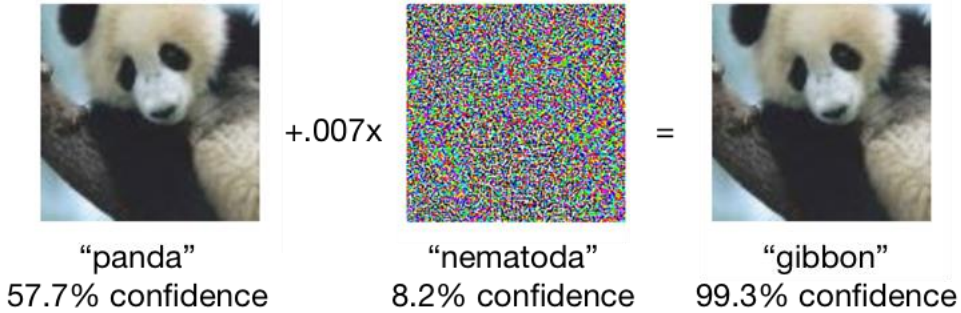


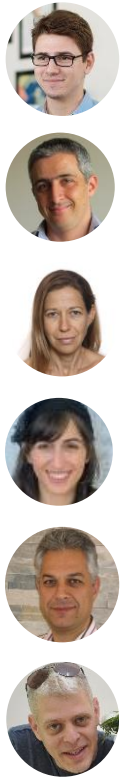
Adversarial: speaker 23



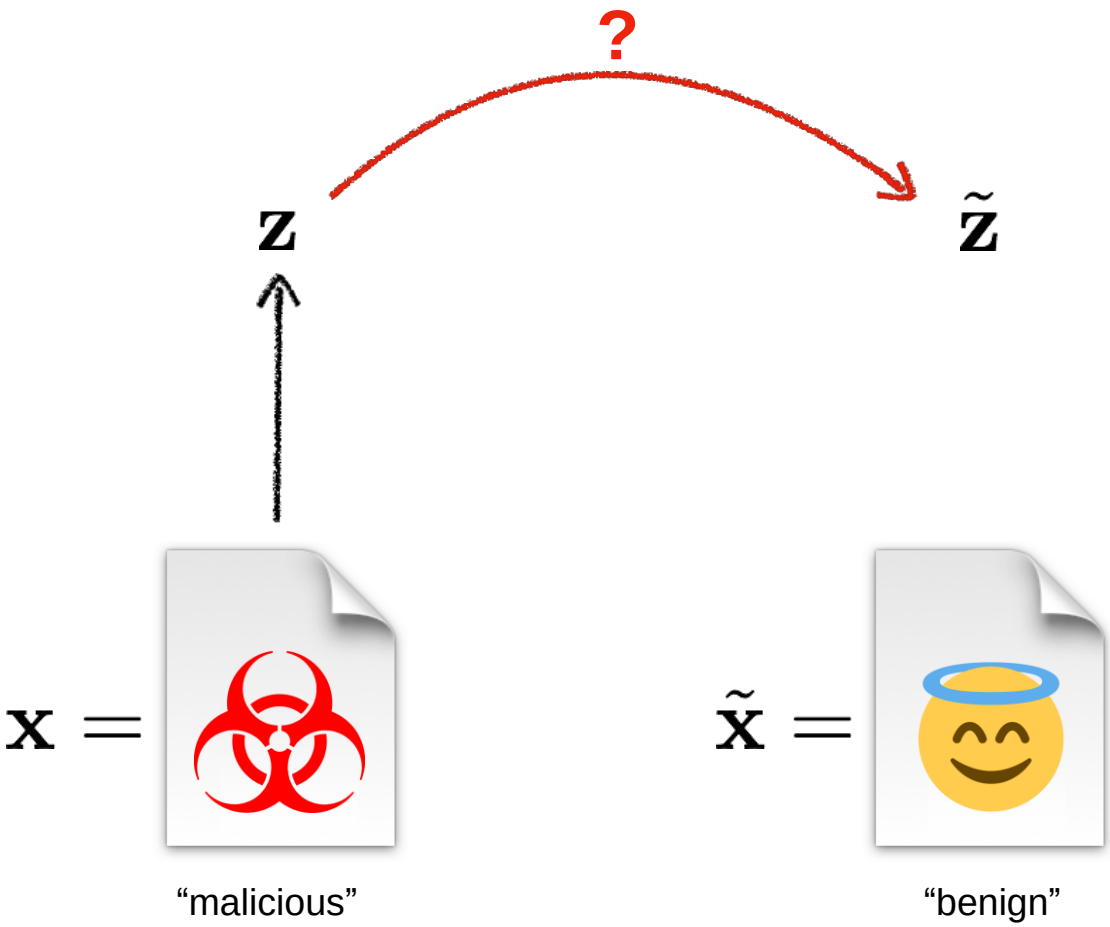
Kreuk, Adi, Cisse, & Keshet, (2017)

Adversarial Malware





Kreuk, Barak, Aviv-Reuven, Baruch, Pinkas & Keshet (2018)


?
$\mathbf{z}$
$\tilde{\mathbf{z}}$
$\mathbf{x} =$
$\tilde{\mathbf{x}} =$
“malicious”
“benign”

$$\tilde{\mathbf{z}} = \mathbf{z} + \delta$$

$\mathbf{z}$ → $\tilde{\mathbf{z}}$

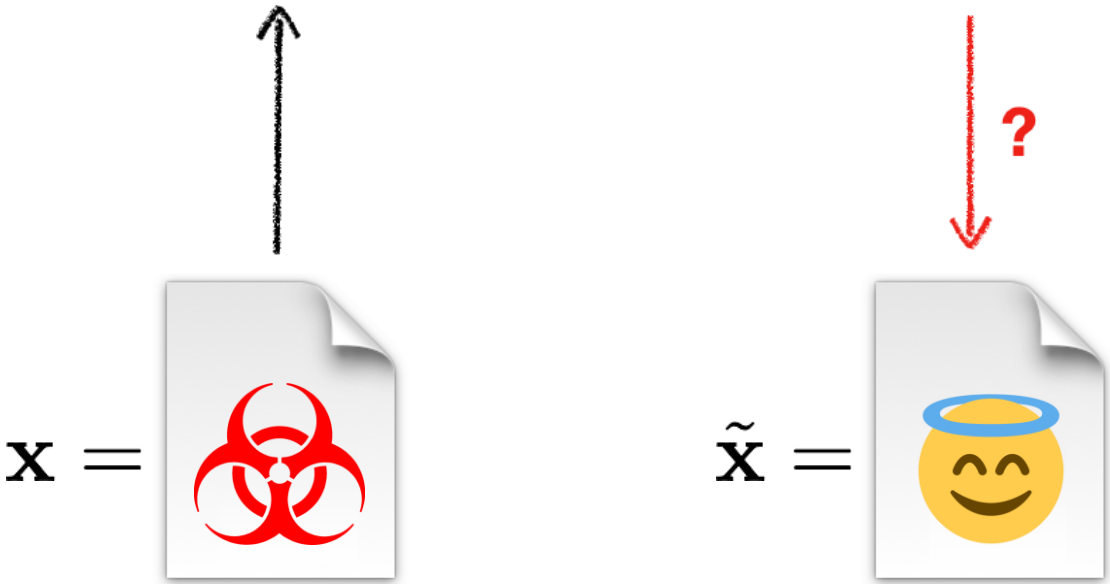
$\mathbf{x} =$ 

$\tilde{\mathbf{x}} =$ 

“malicious”

“benign”

$$\begin{matrix}\mathbf{z}_1 \\ \mathbf{z}_2 \\ \\ \mathbf{z}_L\end{matrix}\begin{bmatrix}0.78 & 0.33 & \dots & 0.13 \\ 0.21 & 0.91 & \dots & 0.55 \\ \dots & \dots & \dots & \dots \\ 0.80 & 0.22 & \dots & 0.63\end{bmatrix} + \begin{bmatrix}0.08 & 0.03 & \dots & 0.09 \\ 0.01 & 0.01 & \dots & 0.04 \\ \dots & \dots & \dots & \dots \\ 0.09 & 0.02 & \dots & 0.01\end{bmatrix} = \begin{bmatrix}0.86 & 0.36 & \dots & 0.22 \\ 0.22 & 0.92 & \dots & 0.59 \\ \dots & \dots & \dots & \dots \\ 0.89 & 0.24 & \dots & 0.64\end{bmatrix}$$



“malicious”

“benign”

Nearest Neighbour Decoding

$$\mathbf{z}_i = \begin{bmatrix} 0.86 & 0.36 & \dots & 0.22 \end{bmatrix}$$

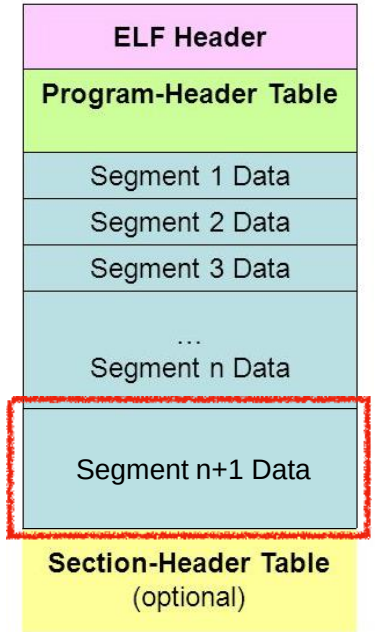
$$\mathbf{M} = \begin{array}{c} \begin{matrix} \mathbf{0} & \mathbf{1} & \mathbf{\dots} & \mathbf{7} \end{matrix} \\ \begin{bmatrix} 0.09 & 0.17 & \dots & 0.55 \\ 0.69 & 0.29 & \dots & 0.69 \\ 0.66 & 0.99 & \dots & 0.08 \\ 0.51 & 0.52 & \dots & 0.24 \\ 0.72 & 0.99 & \dots & 0.93 \\ 0.78 & 0.33 & \dots & 0.13 \\ 0.21 & 0.91 & \dots & 0.55 \\ \dots & \dots & \dots & \dots \\ 0.80 & 0.22 & \dots & 0.63 \end{bmatrix} \end{array} \begin{matrix} \mathbf{0} \\ \mathbf{1} \\ \mathbf{2} \\ \mathbf{3} \\ \mathbf{4} \\ \mathbf{5} \\ \mathbf{6} \\ \mathbf{\dots} \\ \mathbf{255} \end{matrix}$$

Nearest neighbour is byte 5!

Better Decoding

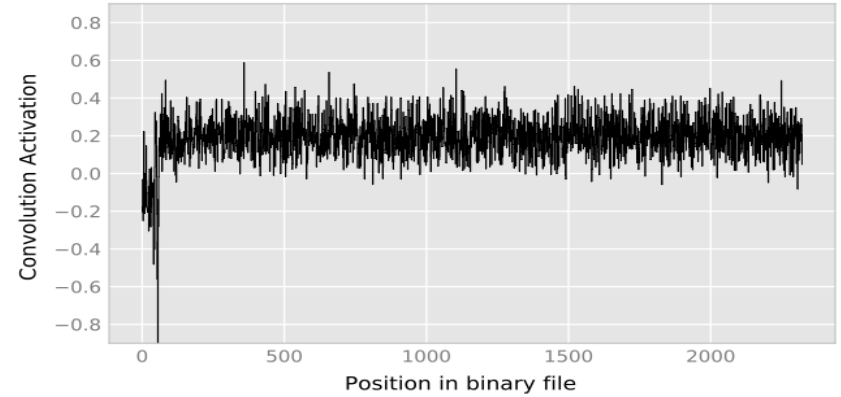
$$\bar{\ell}^*(\mathbf{z}, y; \theta) = (1 - \alpha)\, \bar{\ell}(\mathbf{z}, y; \theta) + \alpha \sum_{i=1}^{L} \min_j d(\mathbf{z}_i, \mathbf{M}_j)$$

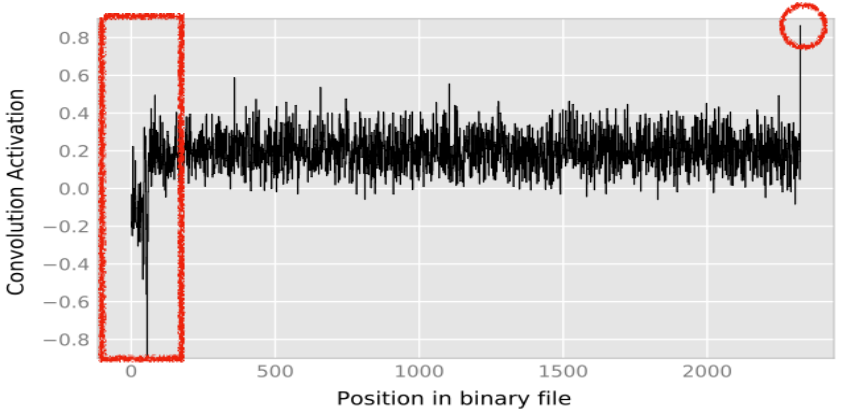
Are We There Yet?



Nearest Neighbor Decoding

p-norm	Evasion rate	Benign confidence
$p = 2$	88%	0.86
$p = \infty$	100%	0.99


Convolution Activation
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
0
500
1000
1500
2000
Position in binary file


Convolution Activation
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
0
500
1000
1500
2000
Position in binary file


No real defense for now…
CNN CNN CNN

Turning your weakness into a strength

Machine learning as a Service

Labeled data

datoin

amazon web services™

ForecastThis

bigml®

Google Cloud Platform

MetaMind

Microsoft Azure

Deep Cognition

Machine learning model

Watermark

Watermarking Deep Neural Networks by Backdooring

training set

training

adapting

“pre-trained” model

trigger set

training set and trigger set

training

“from scratch” model

Adi, Baum, Cisse, Pinkas, and Keshet (2018)

Watermarking Deep Neural Networks by Backdooring

Functionality-preserving a model with a watermark is as accurate as a model without it.

Non-trivial ownership an adversary is not able to claim ownership of the model also if he knows the watermarking algorithm.

Unforgeability an adversary, even when possessing several trigger set examples and their targets, will not be unable to convince a third party about ownership.

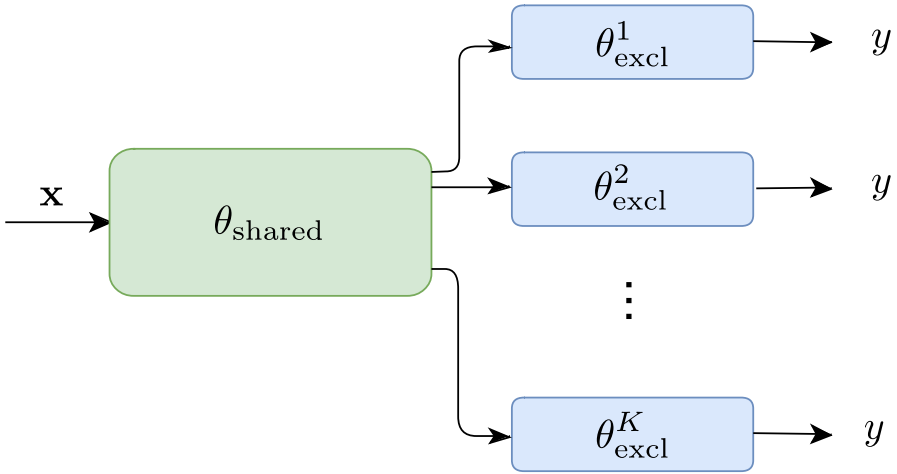
Unremovability an adversary is not able to remove a watermark, even if he knows about the existence of a watermark.

Adi, Baum, Cisse, Pinkas, and Keshet (2018)

Detecting adversarial attacks by adding redundancies

“computer keyboard”

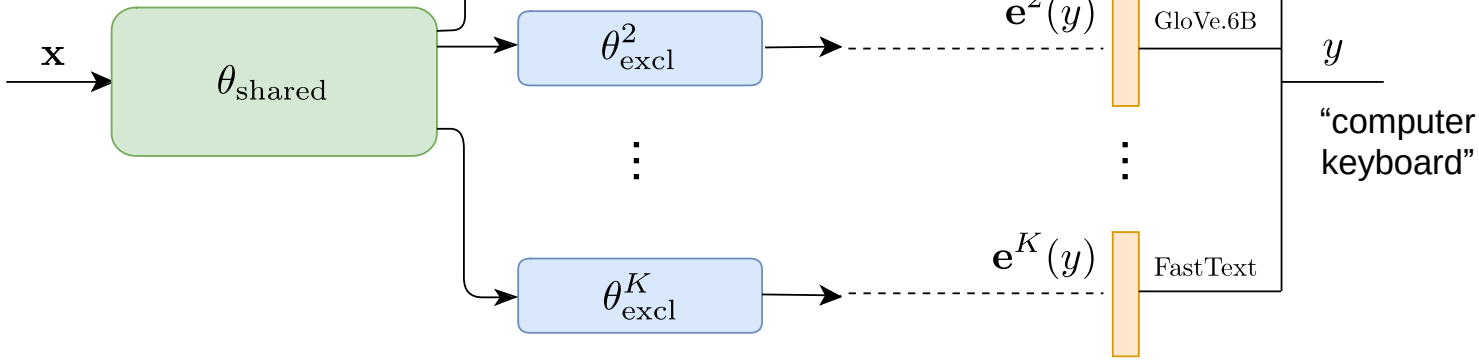
Shalev, Adi, and Keshet (2018)



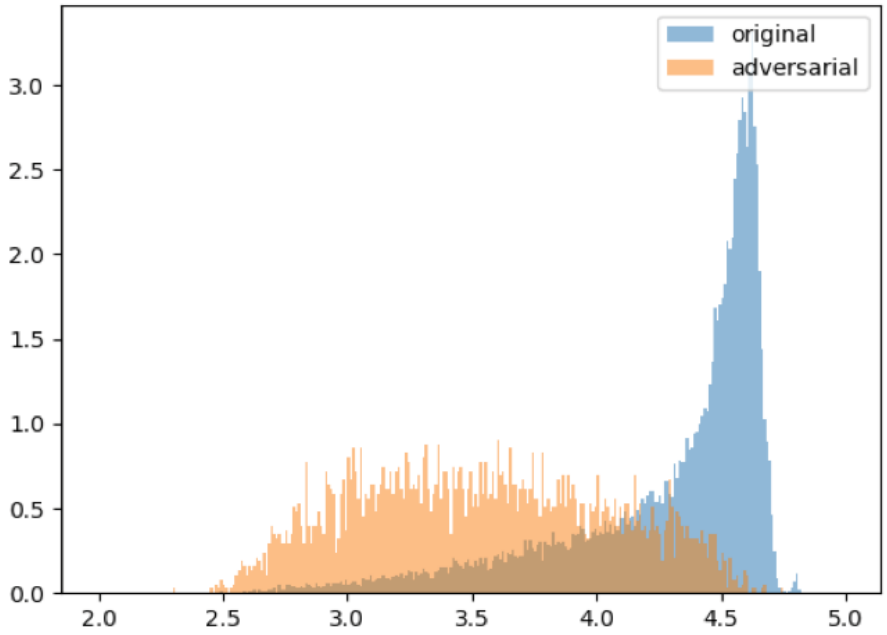
“computer keyboard”

Shalev, Adi, and Keshet (2018)

different vector
representations of words



Shalev, Adi, and Keshet (2018)



Shalev, Adi, and Keshet (2018)

different vector
representations of words

$\mathbf{x}$

θ_{shared}

θ^1_{excl}

θ^2_{excl}

θ^K_{excl}

$\mathbf{e}^1(y)$

$\mathbf{e}^2(y)$

$\mathbf{e}^K(y)$

Error correcting code

y

“computer keyboard”

Keshet, Aviv-Reuven, and NEC (under submission)

different vector
representations of words

$\mathbf{x}$

θ_{shared}

θ^1_{excl}

θ^2_{excl}

θ^K_{excl}

$\mathbf{e}^1(y)$

$\mathbf{e}^2(y)$

$\mathbf{e}^K(y)$

Error correcting code

y

“computer
keyboard”

cryptographic key

cryptographic key

Keshet, Aviv-Reuven, and NEC (under submission)

Speech, Language and Deep Learning Lab


Yossi Adi


Felix Kreuk


Tzevyia Fuchs


Shir Aviv


Gabi Shalev

facebook research


Moustapha Cisse


Natalia Neverova

Center for Research in Applied Cryptography and Cyber Security

BIU


Assi Barak


Carsten Baum


Benny Pinkas


Morna Baruch

Thanks !!