Clarifying the power and limitation of CyberAttacks with Adversarial Examples

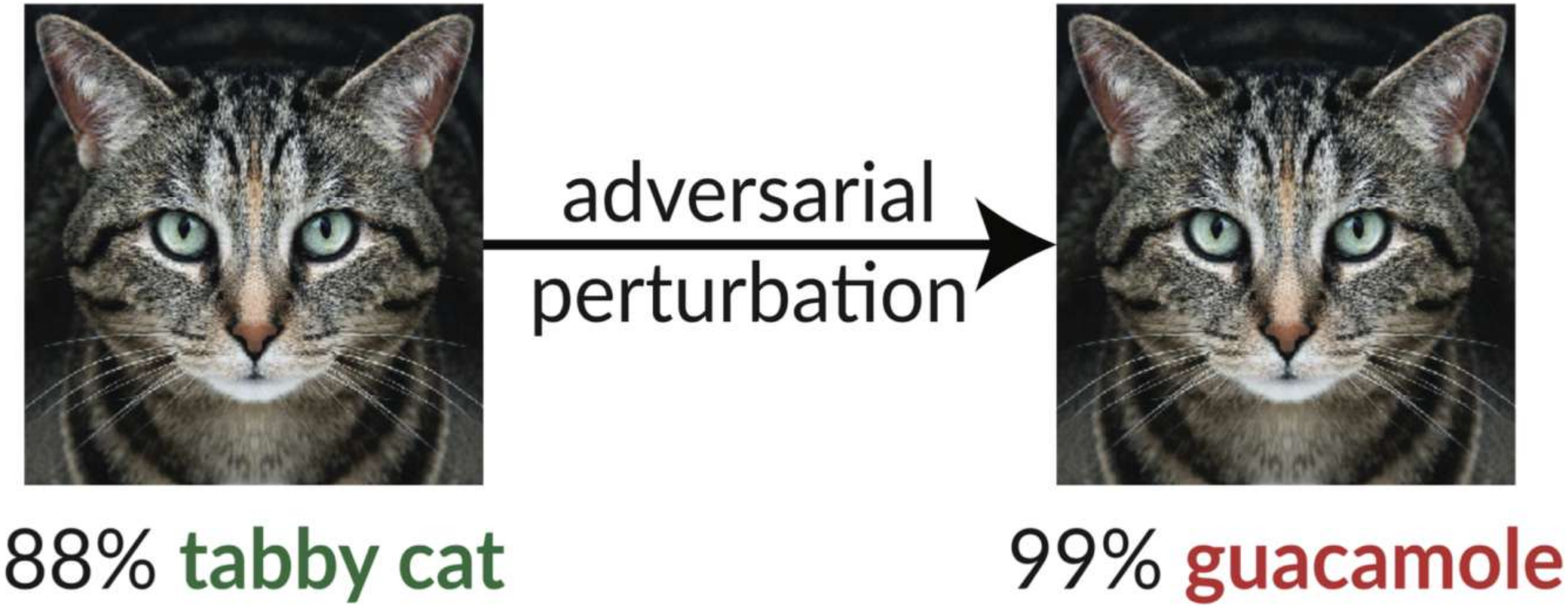


Fig.1. A small change imperceptible to humans misleads the InceptionV3 net- work into classifying an image of a tabby cat as guacamole. Image taken from https://github.com/anishathalye/obfuscated-gradients.

WAIS2020 Feb. 21st
Clarifying on the power and limitation of CyberAttacks with Adversarial Examples

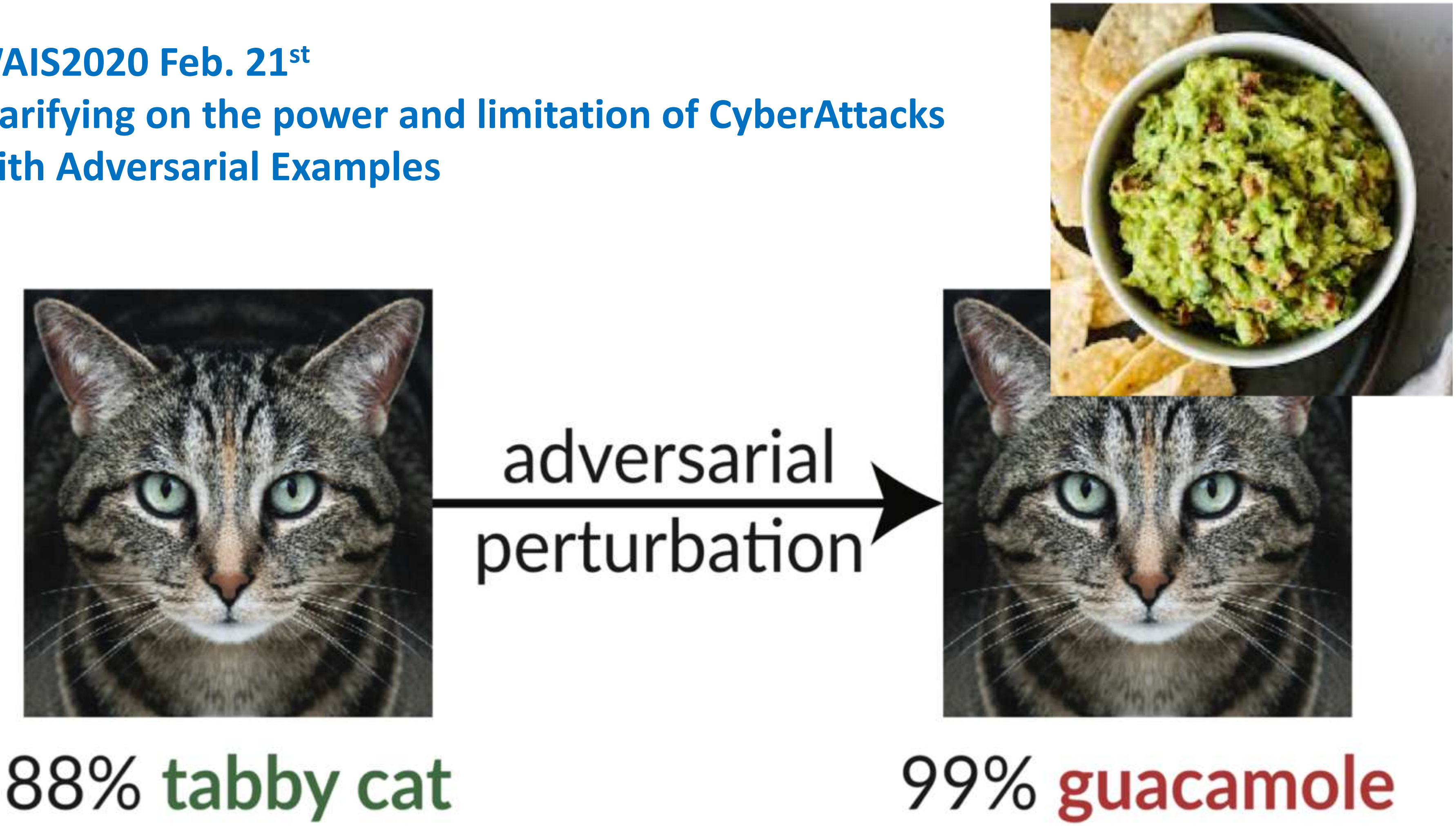


Fig.1. A small change imperceptible to humans misleads the InceptionV3 net- work into classifying an image of a tabby cat as guacamole. Image taken from https://github.com/anishathalye/obfuscated-gradients.

WAIS2020 東京理科大　葛飾　FEB.21ST

Clarifying the power and limitation of CyberAttacks with Adversarial Examples

Adi Shamir Join AI-Research 2019
With Combinatorial Geometry to discuss the power of machine learning

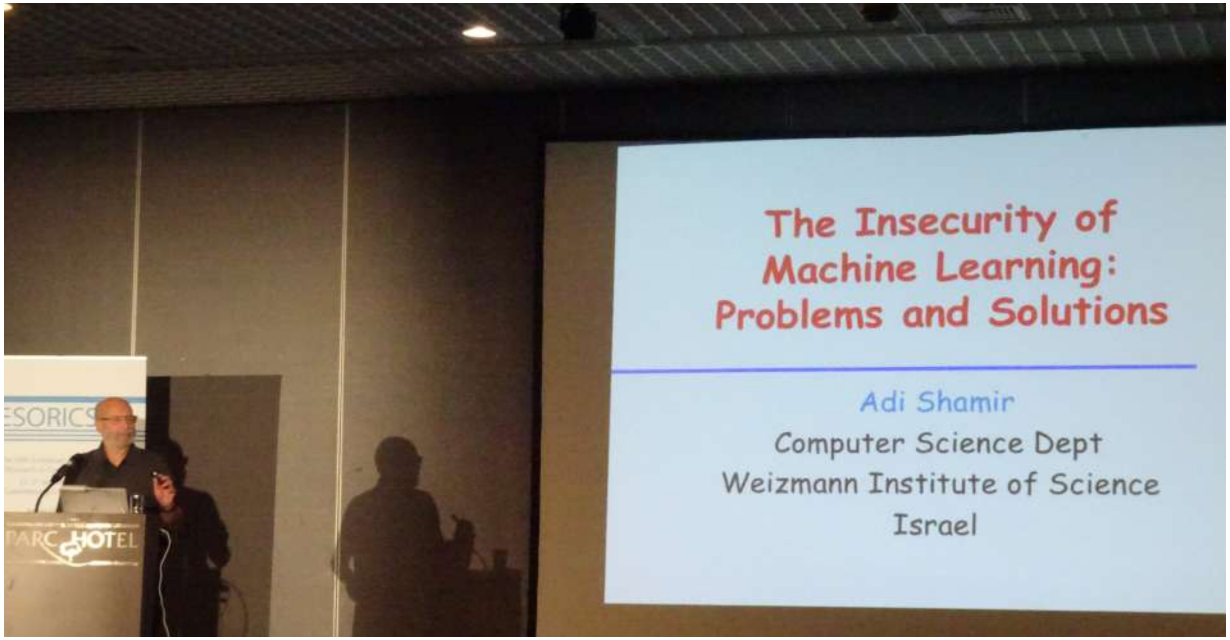
Kouichi SAKURAI 櫻井　幸一
九大：情報学部門&サイバーセキュリティセンター
ATR:　適応コミュニニケーション研究所　先端セキュリティ研究室

Background
It is time for cryptographers 😖
and cybersecurity-researchers 😵
to do study AI 🤖

- Today's talk
  - Introduction : Shamir's recent approach to clarification of adversarial example attack
  - **"One pixel attack for fooling deep neural networks"**
    - My recent great success result

Artificial Intelligence and Cognitive Engineering (ALICE)
University of Groningen
2020 01 30

The Insecurity of Machine Learning: Problems and Solutions


The Insecurity of Machine Learning: Problems and Solutions

Adi Shamir

Computer Science Dept

Weizmann Institute of Science

Israel

1.ESORICS2019

One of Three KeyNotes

- Sept.23rd
- Luxembourg
- PC-chair by Sako(NEC)

2.理研　Center for AIP

Special Lecture

- 2019. 12月3日
- 日本橋

0. arxiv.org/abs/1901.10861

(Submitted on 30 Jan 2019)

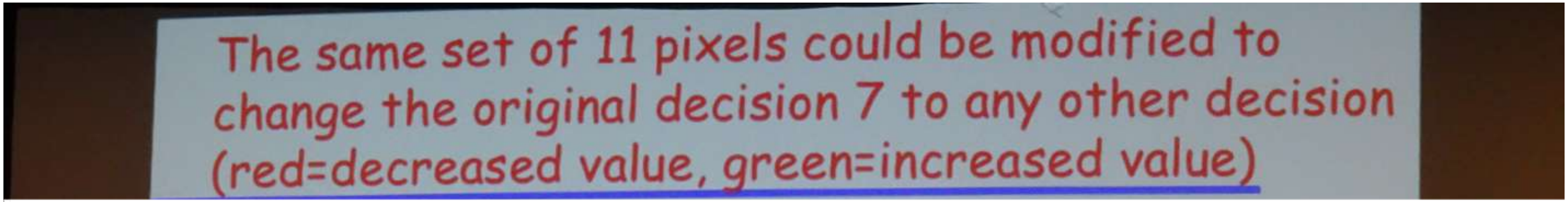
Simple Explanation for the Existence of Adversarial Examples with Small Hamming Distance

Adi Shamir, Itay Safran, Eyal Ronen, Orr Dunkelman

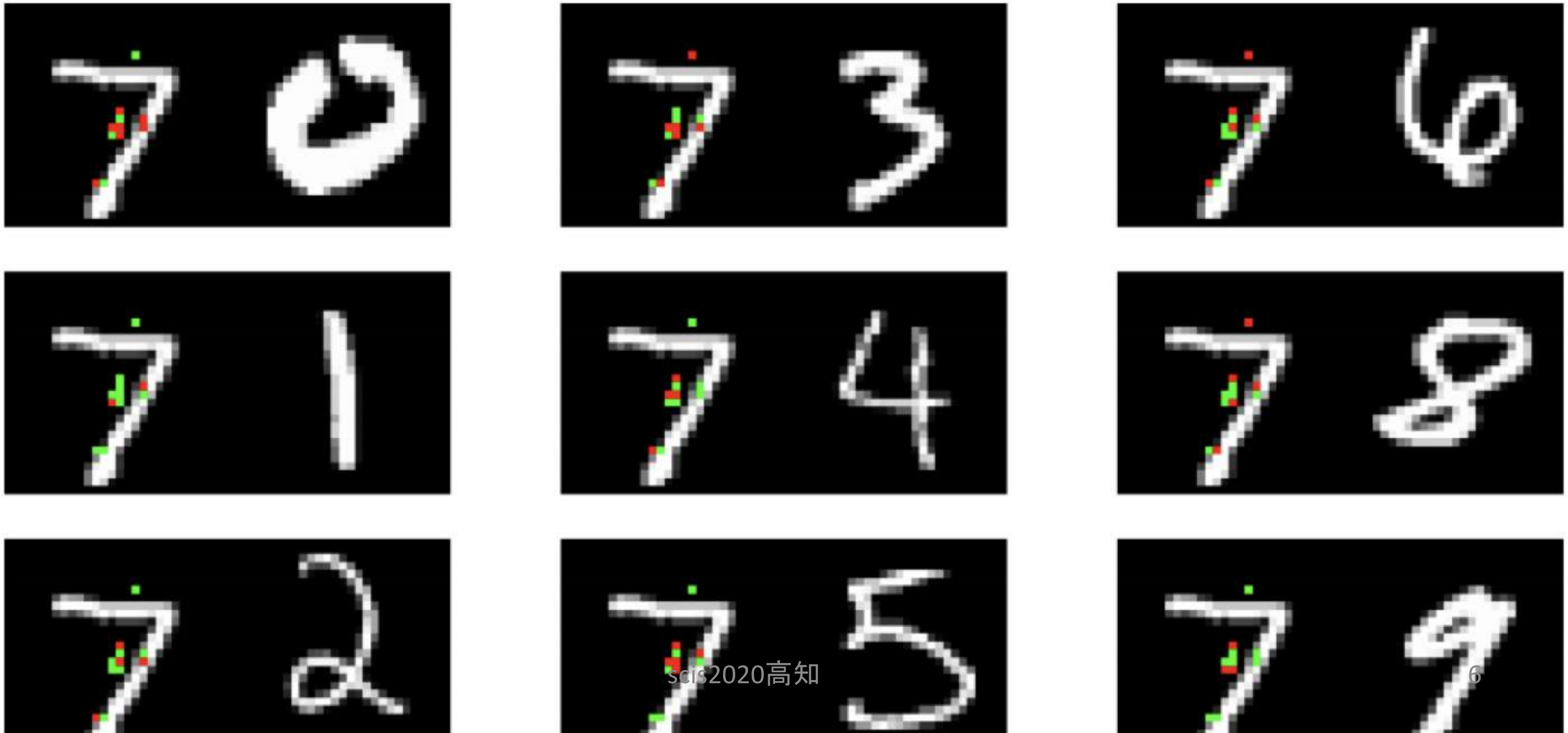
!! 著者順に注意を!! @google-scholar 引用はまだ9件

Shamir’s targeted attacks

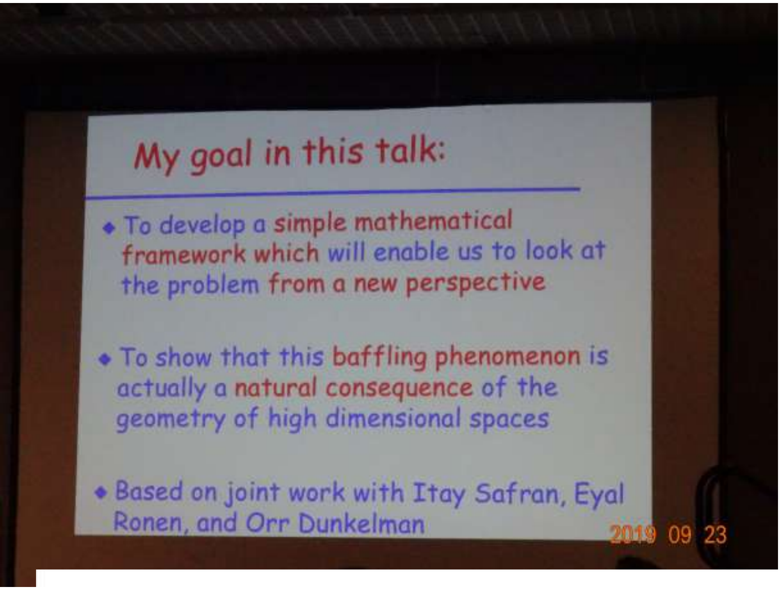
[１１pixel 100% succes but ”10”pixel fail]



The same set of 11 pixels could be modified to change the original decision 7 to any other decision (red=decreased value, green=increased value)



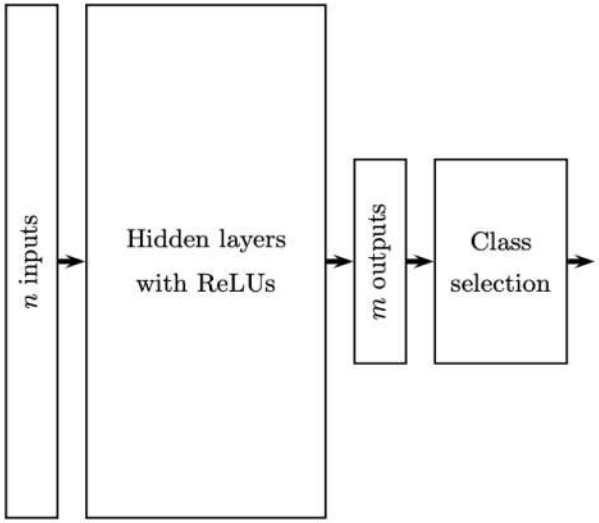
Machine Learning with Big Data@Adi. SHAMIR



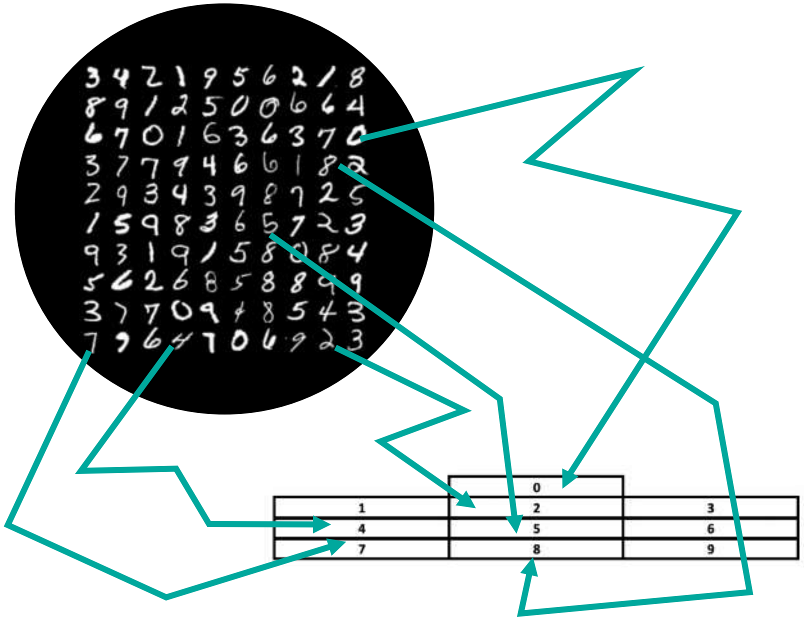
My goal in this talk:

- To develop a simple mathematical framework which will enable us to look at the problem from a new perspective
- To show that this baffling phenomenon is actually a natural consequence of the geometry of high dimensional spaces
- Based on joint work with Itay Safran, Eyal Ronen, and Orr Dunkelman

2019 09 23



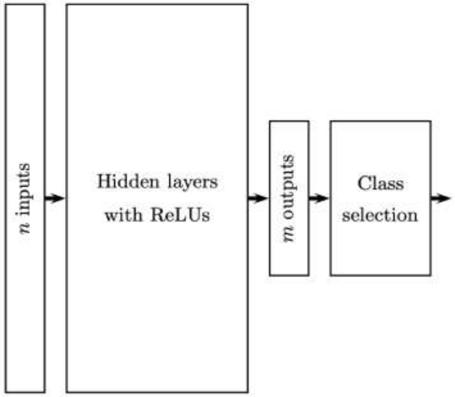
Shamir's Piecewise-Linear map from MNIST
to the the classification : [0,1,2,3,4,5,6,7,8,9]

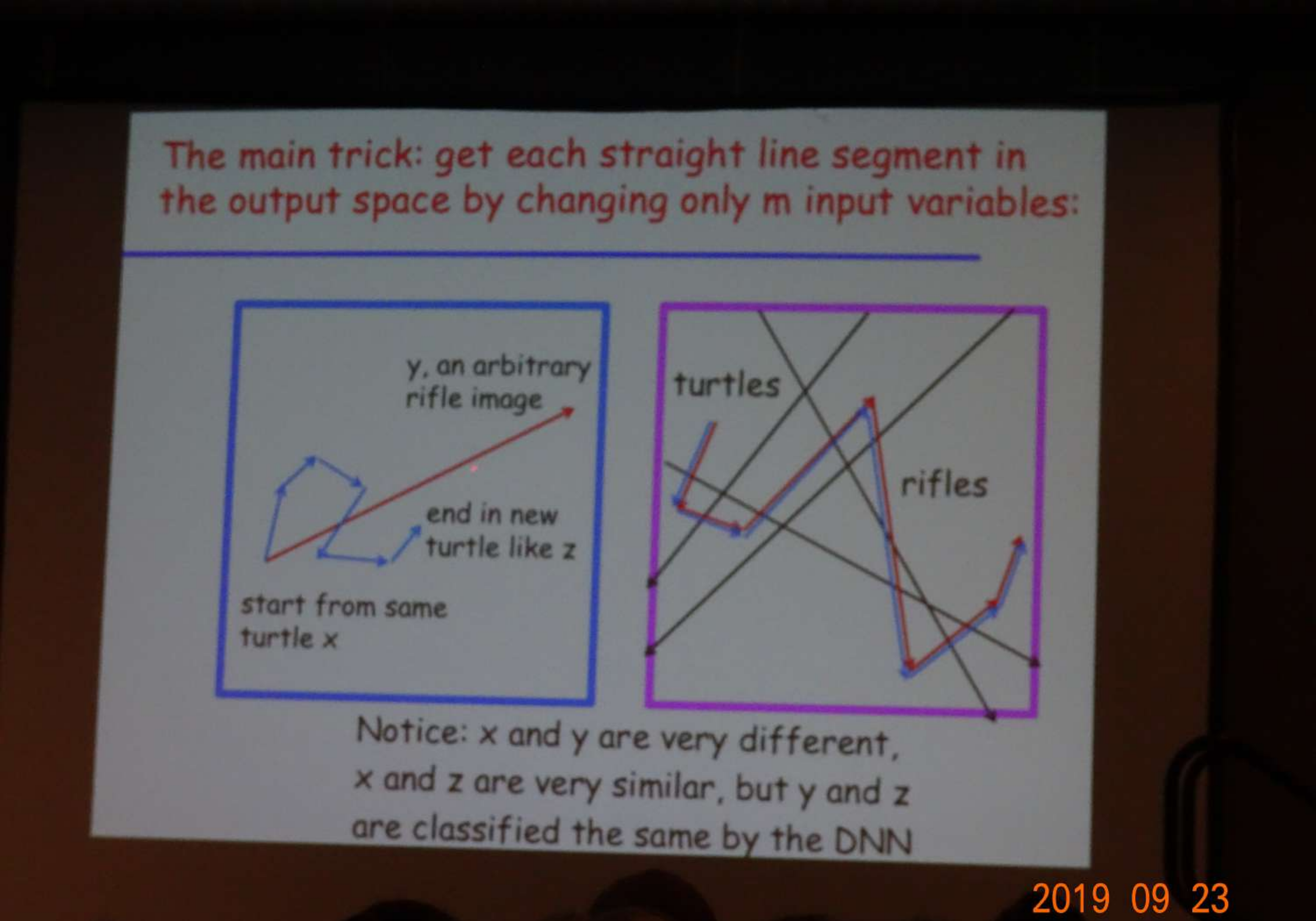


Note: Shamir's experience is the most simple and not exactly deep: ANN with just one hidden layer.

Theory and Experiment by Shamir

- DNN: R^n ➔ R^m :Peicewise Linear (with ReLu)
  - (n>>m): MNIST has (n=784 , m=10)
- L0-norm（Hamming Distance） [L2-case is unsolved]
- Combinatorical Geometry（confirmed Prof. E. Bannai）
  - **T. Zaslavsky, “Facing up to Arrangements·: Face-Count Formulas for Partitions of Space by Hyperplanes”, (1975)**
  - N. Alon, P. Frankl, V. Rodl “GEOMETRICAL REALIZATION OF SET SYSTEMS AND PROBABILISTIC COMMUNICATION COMPLEXITY” (1985)
- Design searching algorithm
  - Swap m-bit ➔ (m+1)-bit
  - Experimental techniques against soft error





The main trick: get each straight line segment in the output space by changing only m input variables:

y, an arbitrary rifle image

end in new turtle like z

start from same turtle x

turtles

rifles

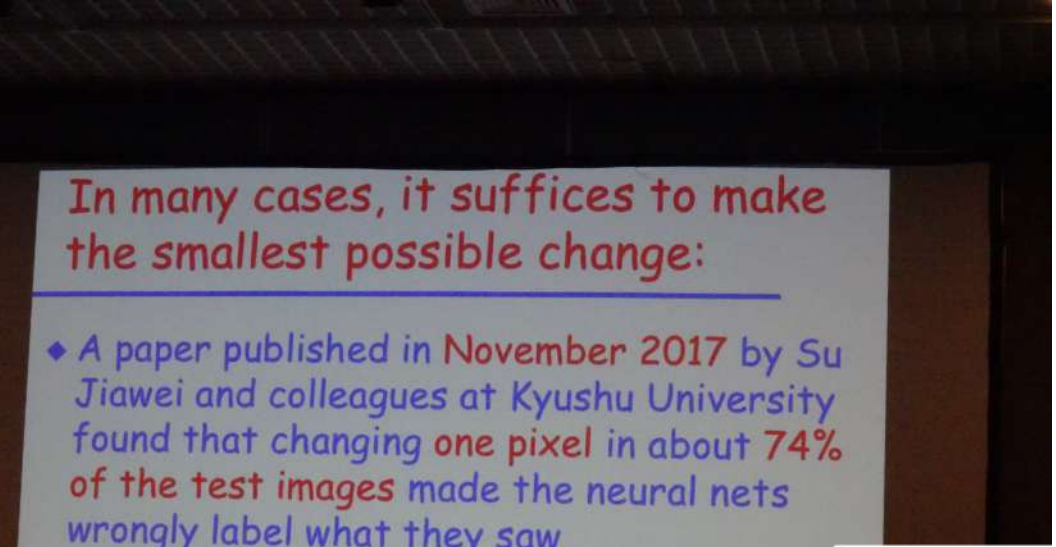
Notice: x and y are very different, x and z are very similar, but y and z are classified the same by the DNN

2019 09 23

Consequences from Shamir

- CONJECTORE： applicable to abt DNN with Peicewise Linear
- Limitation of DNN
  - No use of increasing the number of layer nor complexing the structure, if they are piecewise linear
- Some networks has non piecewise linear
  - [Osadchy et al. “No Bot Expects the DeepCAPTCHA! Introducing Immutable Adversarial Examples, With Applications to CAPTCHA Generation. IEEE Trans. 2017.]

    medium filter with not continuous
  - If this, Shamir’s attack cannot work….

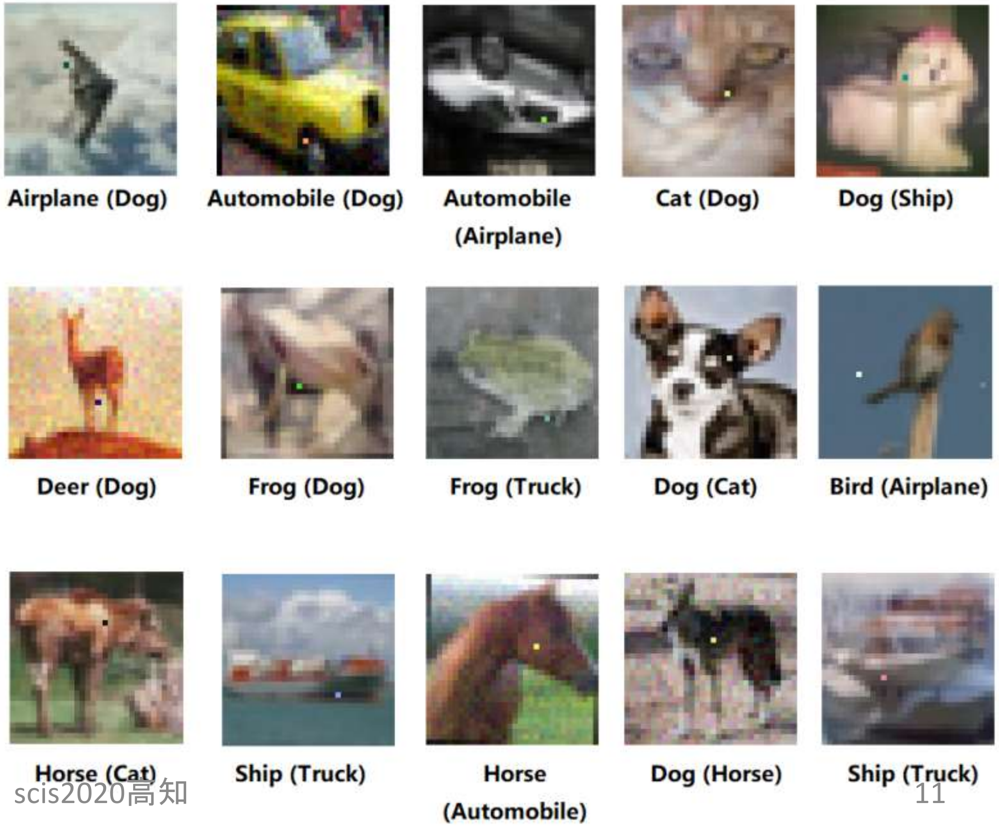


In many cases, it suffices to make the smallest possible change:

- A paper published in November 2017 by Su Jiawei and colleagues at Kyushu University found that changing one pixel in about 74% of the test images made the neural nets wrongly label what they saw

[九大2017.Nov]
蘇(現・KDDI-Lab)-VARGAS-櫻井
One-pixel Attack

Su, J., Vasconcellos Vargas, D., and Kouichi, S. (🡐 氏名が逆)
One pixel attack for fooling deep neural networks.
arXiv:1710.08864, 2017.



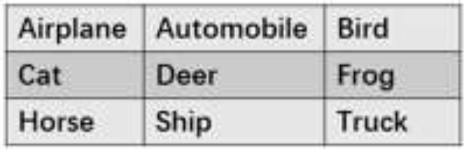
Airplane (Dog) Automobile (Dog) Automobile (Airplane) Cat (Dog) Dog (Ship)

Deer (Dog) Frog (Dog) Frog (Truck) Dog (Cat) Bird (Airplane)

Horse (Cat) Ship (Truck) Horse (Automobile) Dog (Horse) Ship (Truck)

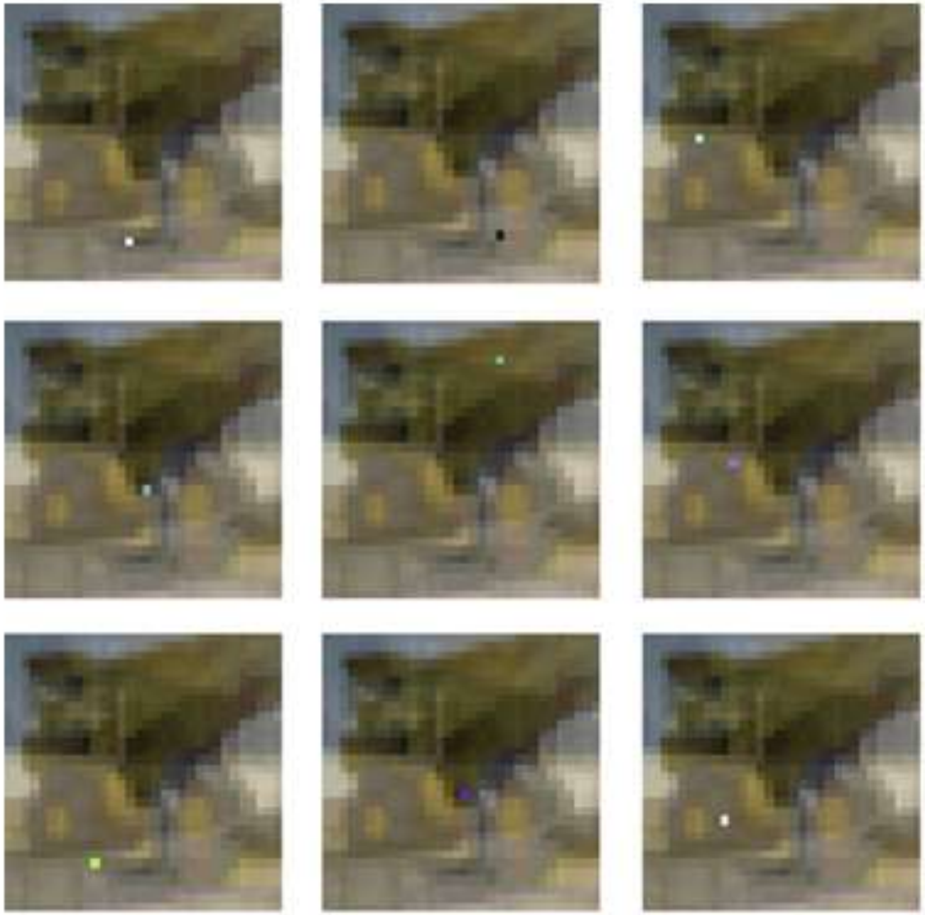
One pixel attack examples(2)

Original image (dog)



Airplane	Automobile	Bird
Cat	Deer	Frog
Horse	Ship	Truck

Target classes



One image can be simultaneously perturbed to nine other classes.

One-Pixel Attack(3)

- 2019. Jan
  - IPSJ Transactions on Computer Vision and Applications
    - ”Attacking convolutional neural network using differential evolution”
  - IEEE Transactions on Evolutionary Computation (Early Access)
    - “One Pixel Attack for Fooling Deep Neural Networks”
    - Cite by **428** papers @google scholar 2020.Jan.30 (→ 447 today)
- **2018 Mar.**
  - **Keynote ICT.OPEN/NL**
- **2018 Feb. Rejected by CVPR**

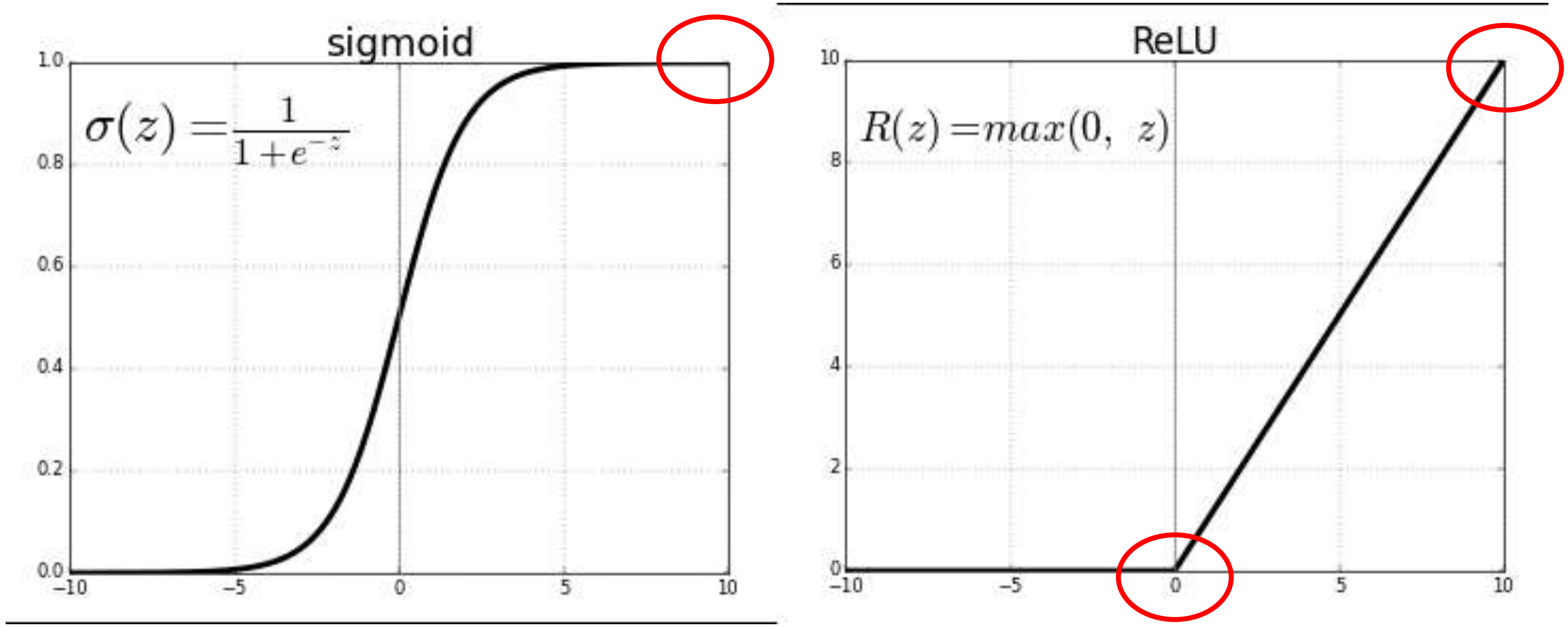
ICT.OPEN [2018 Mar, Amersfoort, NL]

- INVITED “ Power and limitation of Adversarial Machine Learning and their consequences

2018 03 19

- Three Questions
  - One is on why One-Pixel-Attack happen !
    - **From Dr. Sheng HE [贺胜] (Univ. Groningen)**
    - **Because of the characteristic of Activate function !!**
    - **Change ReLU to SIGMOID !!!**

Shapes of Activations (SIGMOID vs ReLU)



Limited

Not Limited

Non-decreasing

J. Su, D. V. Vargas, K. Sakurai
"Empirical Evaluation on Robustness of Deep Convolutional Neural Networks Activation Functions Against Adversarial Perturbation"
CANDAR workshop 2018

Dear Prof. Sakurai Kouichi,

Thanks very much for your nice work and paper.

It looks like that Sigmoid can protect the one-pixel attack, in somehow.

Another solution, might replace the max-pooling with the median-pooling, liking the median filter (https://en.wikipedia.org/wiki/Median_filter).

The position of one-pixel attack and the corresponding wrong class is also very interesting to investigate.

Again, thanks for your invitation for the co-author.

However, I did not do anything about your paper, so I think I am not deserved.

Good luck and best wishes! Sheng

Artificial Intelligence and Cognitive Engineering (ALICE)
University of Groningen

2020 01 30

Professional AI-research ？

- I would-be expert

➔ **Don't just believe it, you have to think about it**

  - Everybody enjoy AI-research！

[GitHAB2019.Jan]
Keras implementation of ”OPA" using differential evolution on Cifar10 and ImageNet

- Open Source
- His result shows our OPA is not so good performance as experimented in our IEEE Trans.

😰

Recent results around OPA

- 2019.Feb. by D.V.Vargas and J.Su
  - Understanding the One-Pixel Attack:

    Propagation Maps and Locality Analysis
- 2019.June, MIT-researchers
  - **“Adversarial Examples Are Not Bugs, TheyAre Features”**

  addresses adversarial examples by removing "non-robust"

    - Features from the training data to only include "human-obvious" features, and shows that algorithms trained on it are more robust against attacks introducing additional features:

It is time for cryptographers
and cybersecurity-researchers
to do study AI

Now with AI-research

Join or Perish

BEYOND YOUR ACCADEMIC DISCIPLINE

- CRYPTOGRAPHY
- CYBERSECURITY
- MATHEMATICS
- ARTIFICIAL INTELIGENCE

Our Happy News from Adi. Shamir

- Mathematics now/easy to go into AI-research
- Combinatorial Geometry
  - Linear Algebra
  - Discrete Probability

  ～APPROXIMATION

Any questions or comments/opinions ?

http://www.escherinhetpaleis.nl/foto

EZWD20772

one-pixel attack(そのI) [蘇-Vargas-櫻井2017]

1. 2017/10/24: Arxivへ論文を投稿し公開発表

   著者：Su, Vargas, and Sakurai

2. 2017/10/30 MIT Tech Review review でArxiv論文が紹介される:

   “How Do You Turn a Dog into a Car? Change a Single Pixel”.

3. 2017/11/02: Su宛にBBC記者から当該論文の関する質問eメール

4. 2017.11/03: BBC e-Newsに掲載される

   “Computers can be fooled into thinking a picture of a taxi is a dog just by changing one pixel, suggests research”

5. 2018/01/10 朝日新聞社より、櫻井宛に、当該論文に関する問い合わせのemailが届き、数回の説明と解説を行う。

6. 2018/01/19 朝日新聞朝刊コラムに研究の引用と下名のコメントが掲載

%未報道: 2018/01/17にも、下名宛にNHK報道局科学文化部から電話で問い合わせあり、emailにて回答。

%% O大の分散計算(Dependability)も知っていた（2018.10）

[GitHAB2019.Jan]
Keras implementation of ”OPA" using differential evolution on Cifar10 and ImageNet

- Open Source
- His result shows our OPA is not so good performance as experimented in our IEEE Trans.