Audio Adversarial Examples

Preliminary Talk

Karla Markert, 07 July 2020

Fraunhofer

AISEC

Outline

Table of Contents

Fraunhofer AISEC

About

About Me

Name	Karla Markert
Department	Cognitive Security Technology
Role	Research assistant
Background	Mathematics, political science and computer science

About

About My Department at AISEC

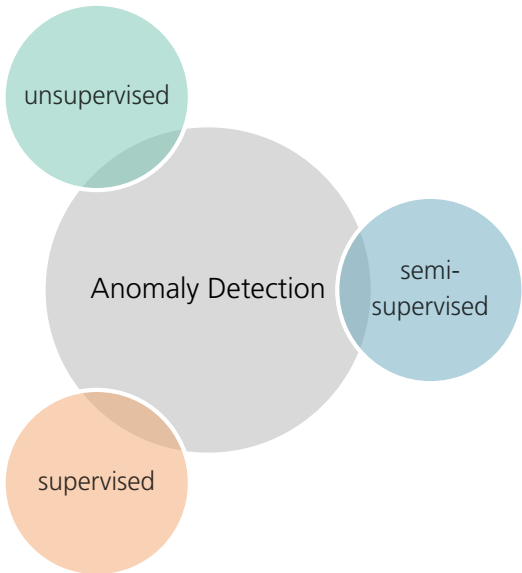
Cognitive Security Technologies:

Intersection of **artificial intelligence** and **IT security**.

Fraunhofer AISEC

About

About My Department at AISEC



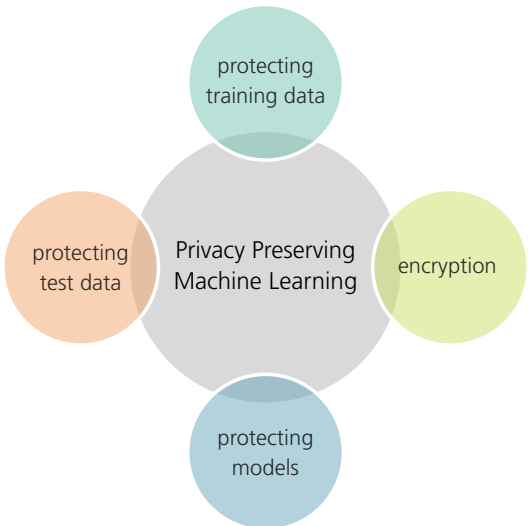
Applications:

- CAN traces,
- malware,
- wireless networks

Fraunhofer AISEC

About

About My Department at AISEC



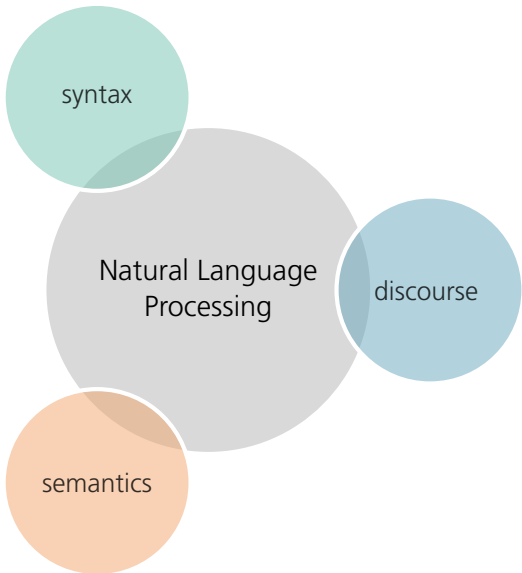
Applications:

- encryption,
- privacy attacks on memory networks,
- architectures for data analysis

Fraunhofer AISEC

About

About My Department at AISEC



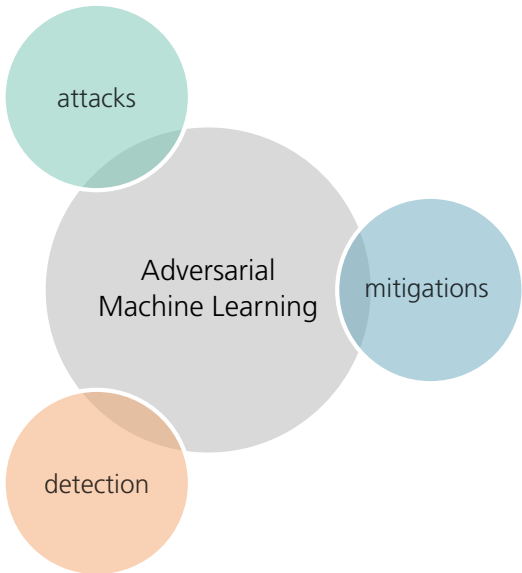
Applications:

- GDPR,
- source code,
- textual descriptions

Fraunhofer AISEC

About

About My Department at AISEC



Applications:

- face recognition,
- speech recognition,
- deep fake detection

Fraunhofer AISEC

Table of Contents

Fraunhofer AISEC

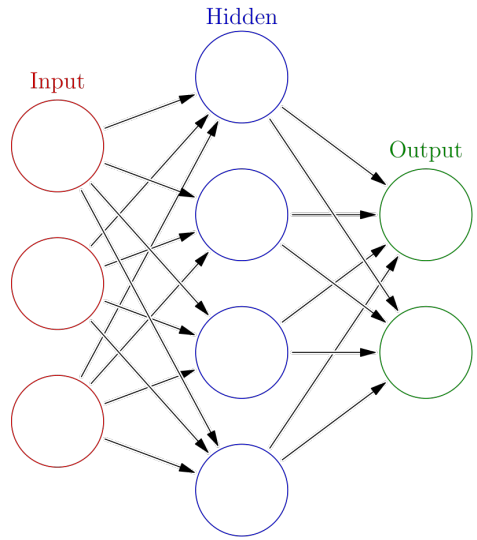
Introduction to Adversarial Examples

Neural Networks

Deep learning "is an approach to AI. Specifically, it is a type of machine learning, a technique that enables computer systems to improve with experience and data. [...] Deep learning is a particular kind of machine learning that achieves great power and flexibility by representing the world as a nested hierarchy of concepts, with each concept defined in relation to simpler concepts, and more abstract representations computed in terms of less abstract ones." [2]

Introduction to Adversarial Examples

Neural Networks



Visualization of a neural network with one hidden layer.

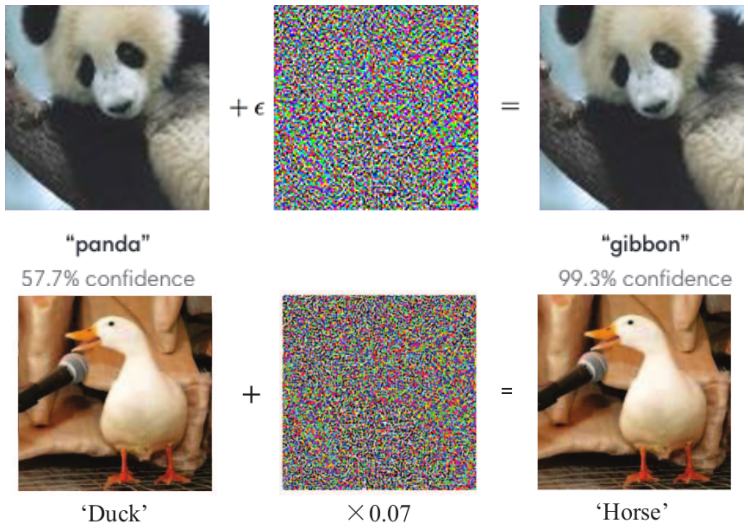
Image taken from Wikipedia[1].

[1] See `https://en.wikipedia.org/wiki/Artificial_neural_network`, last checked January 4, 2020.

Fraunhofer AISEC

Introduction to Adversarial Examples

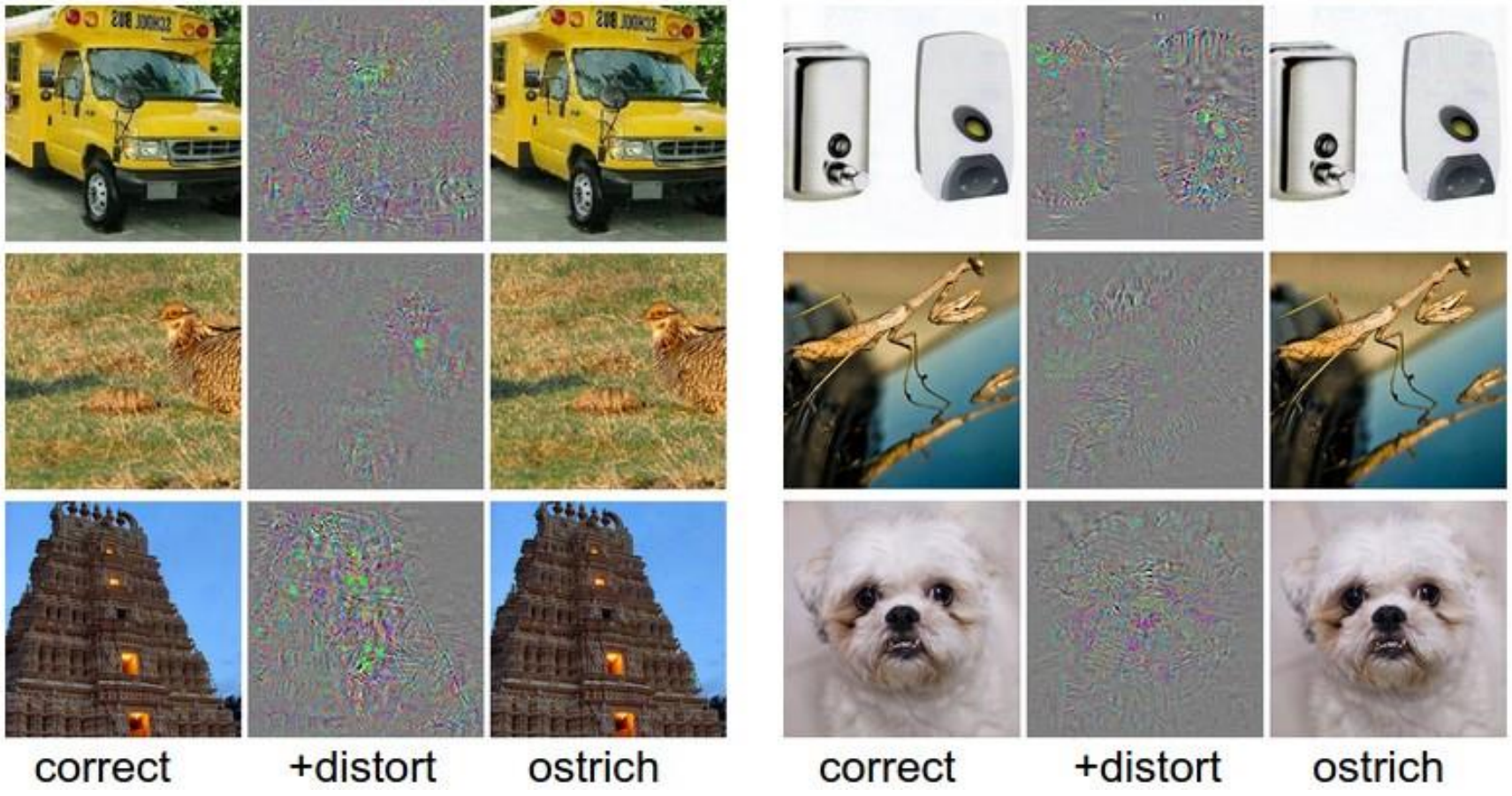
Adversarial Images



Images taken from [5, 4].

Introduction to Adversarial Examples

Adversarial Images



Images taken from [7]. *Ostrich* means *Strauß* in German.

Introduction to Adversarial Examples

Adversarial Images

Stop sign recognized as stop sign.

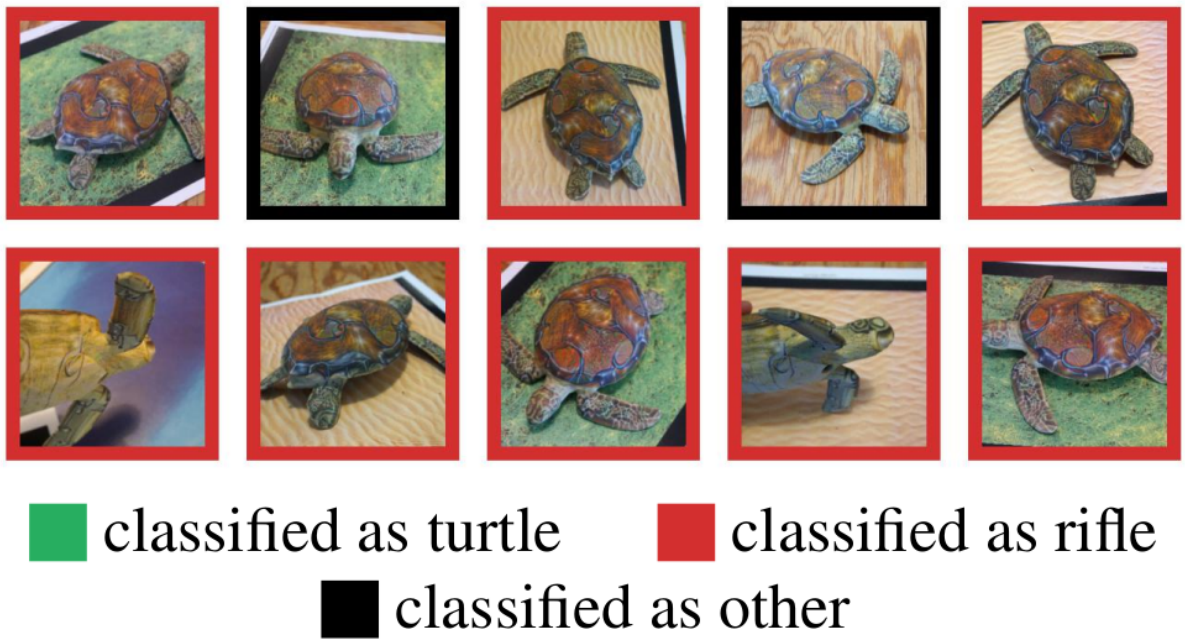
Stop sign recognized as bottles.

Images taken from [6].

Fraunhofer AISEC

Introduction to Adversarial Examples

Adversarial Images



Images taken from [1].

Introduction to Adversarial Examples

Adversarial Audio

Original

Transcription: without the dataset the article is useless

Adversarial

Transcription: okay google browse to evil dot com

Examples taken from [3].

Table of Contents

Fraunhofer AISEC

Organizational Stuff

In this seminar, we take a look at different audio adversarial attacks and possible mitigations.

- **Level:** Bachelor and Master
- **Number of Participants:** 8
- **Language:** English
- **Requirements:** Basic knowledge in machine learning (especially deep neural networks) and IT security.

Organizational Stuff

Time: This course will be held as a block seminar.

- **July 10** (Friday), 14:00 - 14:45 *Preliminary talk*
- **August 11** (Tuesday) 14:00-15:00 *Kick Off*
- **November 26** (Thursday) and **November 27** (Friday), 9:00-17:00 *Presentations*
- **December 4** (Friday), 9:00-10:00 and **January 15** (Friday), 9:00-10:00 *Debriefing*
- **December 7** (Monday), 23:59 *Deadline for paper*

Organizational Stuff

Goals:

- familiarization with scientific paper reading and scientific presentations;
- better understanding of attacks against machine learning algorithms;
- active participation and insights into topics of current research. For more information, see module description IN0014 and IN2107.

Organizational Stuff

Method: The seminar is organized as follows.

- Every participant gives a *presentation on a scientific paper*, which is assigned in the kick off session.
- Every student is required to write a *four page hand out* summarizing the main points of the paper (LaTeX template will be provided).

We attach great importance to all students profiting from the others' presentations.

Organizational Stuff

The **grade** is composed up of:

- 10% active participation,
- 25% presentation (structure of the talk, introduction to the topic, clear problem definition and motivation, sound style of delivery...),
- 25% hand out (language, structure of the hand out...),
- 40% quality of the content (main points of the paper, good discussion and outlook...)

Organizational Stuff

Everyone interested in participating is required to send an e-mail to karla.markert@aisec.fraunhofer.de until July 16 indicating her/his interest. You may include a letter of motivation, a CV or a transcript of records.

...any questions so far?

Bibliography

[1] Anish Athalye et al. "Synthesizing robust adversarial examples". In: *arXiv preprint arXiv:1707.07397* (2017).

[2] Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[3] Nicholas Carlini and David A. Wagner. "Audio Adversarial Examples: Targeted Attacks on Speech-to-Text". In: *CoRR* abs/1801.01944 (2018). arXiv: 1801.01944. URL: http://arxiv.org/abs/1801.01944.

[4] Yuan Gong and Christian Poellabauer. "Protecting voice controlled systems using sound source identification based on acoustic cues". In: *2018 27th International Conference on Computer Communication and Networks (ICCCN)*. IEEE. 2018, pp. 1–9.

[5] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples". In: *arXiv preprint arXiv:1412.6572* (2014).

[6] Jiajun Lu, Hussein Sibai, and Evan Fabry. "Adversarial Examples that Fool Detectors". In: *CoRR* abs/1712.02494 (2017). arXiv: 1712.02494. URL: http://arxiv.org/abs/1712.02494.

[7] Christian Szegedy et al. "Intriguing properties of neural networks". In: *arXiv preprint arXiv:1312.6199* (2013).

Contact Information

Karla Markert

Department
Cognitive Security Technologies

Fraunhofer-Institute for
Applied and Integrated Security (AISEC)

Address: Lichtenbergstr. 11
85748 Garching (near Munich)
Germany
Internet: www.aisec.fraunhofer.de

Phone: +49 89 3229986-136
E-Mail: karla.markert@aisec.fraunhofer.de

Fraunhofer AISEC