Malicious Input Detection for Deep Neural Networks

Emilio R. Balda

Institute for Theoretical Information Technology
RWTH AACHEN UNIVERSITY

Outline

- Personal Background
  - home country
  - bachelor studies
- The Ilmenau Experience
  - life as a MSCSP student in Ilmenau
  - a tensor-based master thesis
- From Ilmenau to Aachen
  - the Institute for Theoretical Information Technology
  - our research fields
- Technical Talk
  - introduction to neural networks
  - malicious input detection
  - results
- Personal advice

Outline

- **Personal Background**
  - home country
  - bachelor studies
- The Ilmenau Experience
  - life as a MSCSP student in Ilmenau
  - a tensor-based master thesis
- From Ilmenau to Aachen
  - the Institute for Theoretical Information Technology
  - our research fields
- Technical Talk
  - introduction to neural networks
  - malicious input detection
  - results
- Personal advice

Personal Background
Home Country

- My full name is Emilio Rafael Balda Cañizares
- I was born in the city of Guayaquil, located on the coast of Ecuador

Ecuador

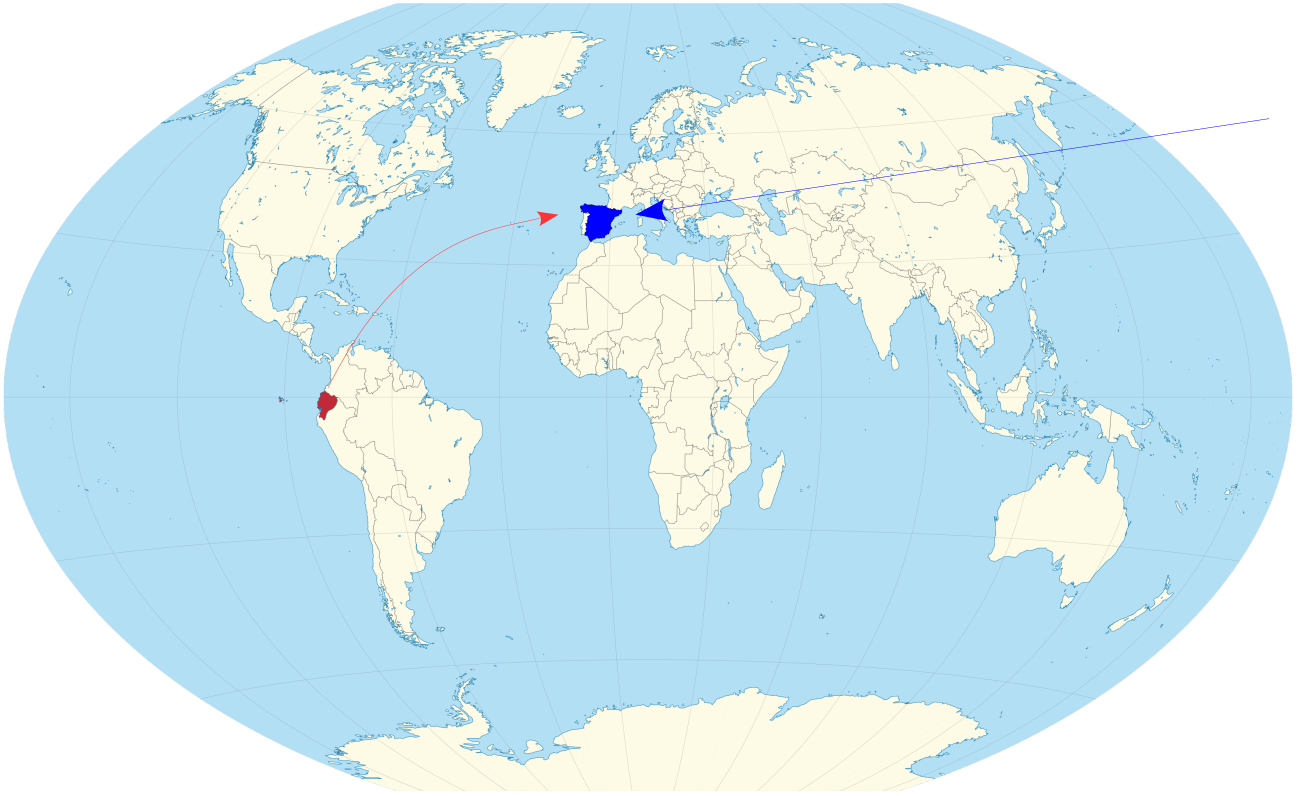
Language:	Spanish
Population:	16 Million
Area:	283,560 km^2

Guayaquil

Elevation:	0 mts
Temperature:	25 to 30 °C
Humidity:	65 to 95%

Personal Background
Bachelor Studies

- Bachelor in Telecommunication Systems Engineering
  - at University of Navarra, Spain
  - in the Engineering faculty known as TECNUN



tecnun
Universidad de Navarra

Outline

- Personal Background
  - home country
  - bachelor studies
- **The Ilmenau Experience**
  - life as a MSCSP student in Ilmenau
  - a tensor-based master thesis
- From Ilmenau to Aachen
  - the Institute for Theoretical Information Technology
  - our research fields
- Technical Talk
  - introduction to neural networks
  - malicious input detection
  - results
- Personal advice

The Ilmenau Experience

Life as a MSCSP Student in Ilmenau

- Personal impression of the MSCSP program
  - strongly research oriented, the perfect choice for future Ph.D. candidates
  - highly specialized on communication networks and signal processing
  - provides several opportunities to explore and be creative on different research topics

TECHNISCHE UNIVERSITÄT ILMENAU

The Ilmenau Experience
Life as a MSCSP Student in Ilmenau

- Personal impression of the MSCSP program
  - strongly research oriented, the perfect choice for future Ph.D. candidates
  - highly specialized on communication networks and signal processing
  - provides several opportunities to explore and be creative on different research topics

TECHNISCHE UNIVERSITÄT ILMENAU

- Life outside the academic
  - large variety of activities and clubs
  - Its easy to adopt the comfortable Ilmenau way of life
  - a 2 years stay allows you to experience many of the activities that Ilmenau offers

The Ilmenau Experience
Life as a MSCSP Student in Ilmenau

- Personal impression of the MSCSP program
  - strongly research oriented, the perfect choice for future Ph.D. candidates
  - highly specialized on communication networks and signal processing
  - provides several opportunities to explore and be creative on different research topics

TECHNISCHE UNIVERSITÄT ILMENAU

- Life outside the academic
  - large variety of activities and clubs
  - Its easy to adopt the comfortable Ilmenau way of life
  - a 2 years stay allows you to experience many of the activities that Ilmenau offers

- I personally can say that I really enjoyed my studies in Ilmenau

The Ilmenau Experience
A Tensor-based Master Thesis

- Master Thesis Title: "*Perturbation analysis of tensor-based algorithms*" - Advisor: Prof. Dr.-Ing. Martin Haardt

The Ilmenau Experience
A Tensor-based Master Thesis

- Master Thesis Title: "*Perturbation analysis of tensor-based algorithms*" - Advisor: Prof. Dr.-Ing. Martin Haardt
  - Conducted a theoretical perturbation analysis of:
    - the truncated **H**igher-**O**rder **S**ingular **V**alue **D**ecomposition (**HOSVD**), mainly used for dimensionality reduction **[1]**
    - the **J**oint **E**igen-**V**alue **D**ecomposition (**JEVD**), mainly used for data analysis **[2]**
    - the **C**anonical **P**olyadic **D**ecomposition (**CPD**), mainly used for data analysis (conference version currently being written)

[1]	E. R. Balda, S. A. Cheema, J. Steinwandt, M. Haardt, A. Weiss, and A. Yeredor, "First-order perturbation analysis of low-rank tensor approximations based on the truncated HOSVD," in *Proceedings of ASILOMAR*, Nov. 2016
[2]	E. R. Balda, S. A. Cheema, A. Weiss, M. Haardt, and A. Yeredor, "Perturbation Analysis of Joint Eigenvalue Decomposition Algorithms," in *Proceesings of ICASSP*, March. 2017

Outline

- Personal Background
  - home country
  - bachelor studies
- The Ilmenau Experience
  - life as a MSCSP student in Ilmenau
  - a tensor-based master thesis
- **From Ilmenau to Aachen**
  - the Institute for Theoretical Information Technology
  - our research fields
- Technical Talk
  - introduction to neural networks
  - malicious input detection
  - results
- Personal advice

Institute for Theoretical Information Technology
RWTH AACHEN UNIVERSITY

From Ilmenau to Aachen
The Institute for Theoretical Information Technology

- The Institute for **T**heoretical **I**nformation Technology (**TI**)
  - belongs to the RWTH Aachen University
  - as part of the Faculty of Electrical Engineering & Information Technology

From Ilmenau to Aachen
The Institute for Theoretical Information Technology

- The Institute for **T**heoretical **I**nformation Technology (**TI**)
  - belongs to the RWTH Aachen University
  - as part of the Faculty of Electrical Engineering & Information Technology

  - is led by Prof. **Dr. Rudolf Mathar**
    - Head of the Institute
    - Pro-Rector of Research and Structure at RWTH Aachen University

From Ilmenau to Aachen
The Institute for Theoretical Information Technology

- The Institute for **T**heoretical **I**nformation Technology (**TI**)
  - is located in Aachen, Germany

From Ilmenau to Aachen
The Institute for Theoretical Information Technology

- The Institute for **T**heoretical **I**nformation Technology (**TI**)
  - is located in Aachen, Germany
  - on the 3rd floor of the ICT Cubes of the RWTH Aachen Univeristy **[BA+11]**

[BA+11] Böcherer, G., Altenbach, F., Malsbender, M., & Mathar, R. “Writing on the facade of RWTH ICT Cubes: Cost constrained geometric Huffman coding.” in Proceedings of *IEEE Wireless Communication Systems (ISWCS)*, 2011.

From Ilmenau to Aachen
The Institute for Theoretical Information Technology

- Staff:
  - Professor Dr. Rudolf Mathar is head of the institute
  - Prof. Dr-Ing. Anke Schmeink is head of the research group *Information Theory and Systematic Design of Communication Systems*


From Ilmenau to Aachen
The Institute for Theoretical Information Technology

- Staff:
  - Professor Dr. Rudolf Mathar is head of the institute
  - Prof. Dr-Ing. Anke Schmeink is head of the research group *Information Theory and Systematic Design of Communication Systems*


  - 2 + 22 academic staff (April 2017)
  - 4 non-academic staff (April 2017)
  - 5 student assistants (April 2017)

From Ilmenau to Aachen
Our Research Fields

- Information Theory & Communication Theory
- Network Design, Control & Optimization
- OFDM Systems
- Compressed Sensing & Signal Classification
- Planning, Simulation, Evaluation for Energy Grids

From Ilmenau to Aachen
Our Research Fields

- Information Theory & Communication Theory
- Network Design, Control & Optimization
- OFDM Systems
- Compressed Sensing & Signal Classification
- Planning, Simulation, Evaluation for Energy Grids
- **Data Analysis and Deep Learning**
  - Theoretical limits for Deep Learning architectures
  - Applying tensor algebra to machine learning problems (*)
  - Machine learning for signal processing and communication applications

Outline

- Personal Background
  - home country
  - bachelor studies
- The Ilmenau Experience
  - life as a MSCSP student in Ilmenau
  - a tensor-based master thesis
- From Ilmenau to Aachen
  - the Institute for Theoretical Information Technology
  - our research fields
- **Technical Talk**
  - introduction to neural networks
  - malicious input detection
  - results
- Personal advice

Technical Talk
Introduction to Neural Networks

- What the world gives us

$\mathbf{x} \in \mathbb{R}^n$

$\mathbf{y} \in \mathbb{R}^m$

observed input

observed output

$$\mathbf{y} = f(\mathbf{x})$$

unknown function

- How do we approximate $f(\cdot)$?
  - **Naive approach**: approximate it with a polynomial and “train” the polynomial weights. Example for *n*=5 and *m*=1
    - e.g. $\hat{y} = w_1(x_1)^3 + w_2(x_1)^2 x_2 + w_3(x_1)^2 x_3 + \cdots + w_N(x_5)^3$

estimation

these weights are “trained” by minimizing

$$\min_{\mathbf{w}} \sum_{\mathbf{x} \in \mathcal{S}} (y - \hat{y})^2$$

large set of observations

Technical Talk
Introduction to Neural Networks

- What the world gives us

$\mathbf{x} \in \mathbb{R}^n$ → Real World System → $\mathbf{y} \in \mathbb{R}^m$

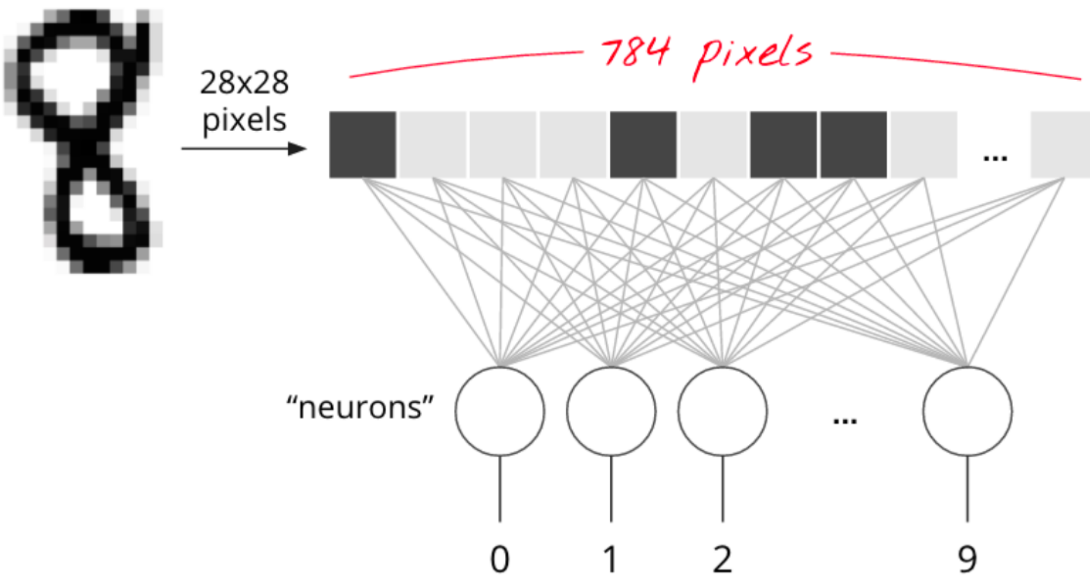
observed input   observed output

$$\mathbf{x} = f(\mathbf{y})$$

unknown function

- How do we approximate $f(\cdot)$?
  - **Better approach**: Use a “neural network” composed of linear and non-linear functions and train its parameters
    - a neural network with finite amount of parameters can approximate a wide variety of functions, see the *universal approximation theorem*
    - cost functions as $\min_{\mathbf{w}} \sum_{\mathbf{x}\in\mathcal{S}} \|\mathbf{y} - \hat{\mathbf{y}}\|_2^2$ are also used here

- Lets start with a neural network of 1 layer



$\mathbf{x} \in \mathbb{R}^n$

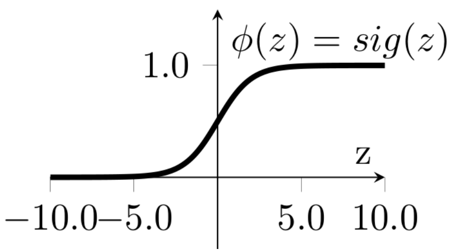
$\mathbf{y} \in \mathbb{R}^m$

Weight matrix

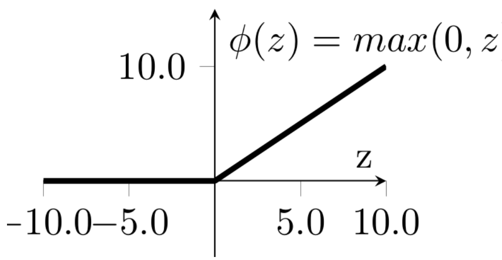
bias vector

$$\mathbf{y} = \phi(\mathbf{W}\mathbf{x} + \mathbf{b})$$

point-wise nonlinear function



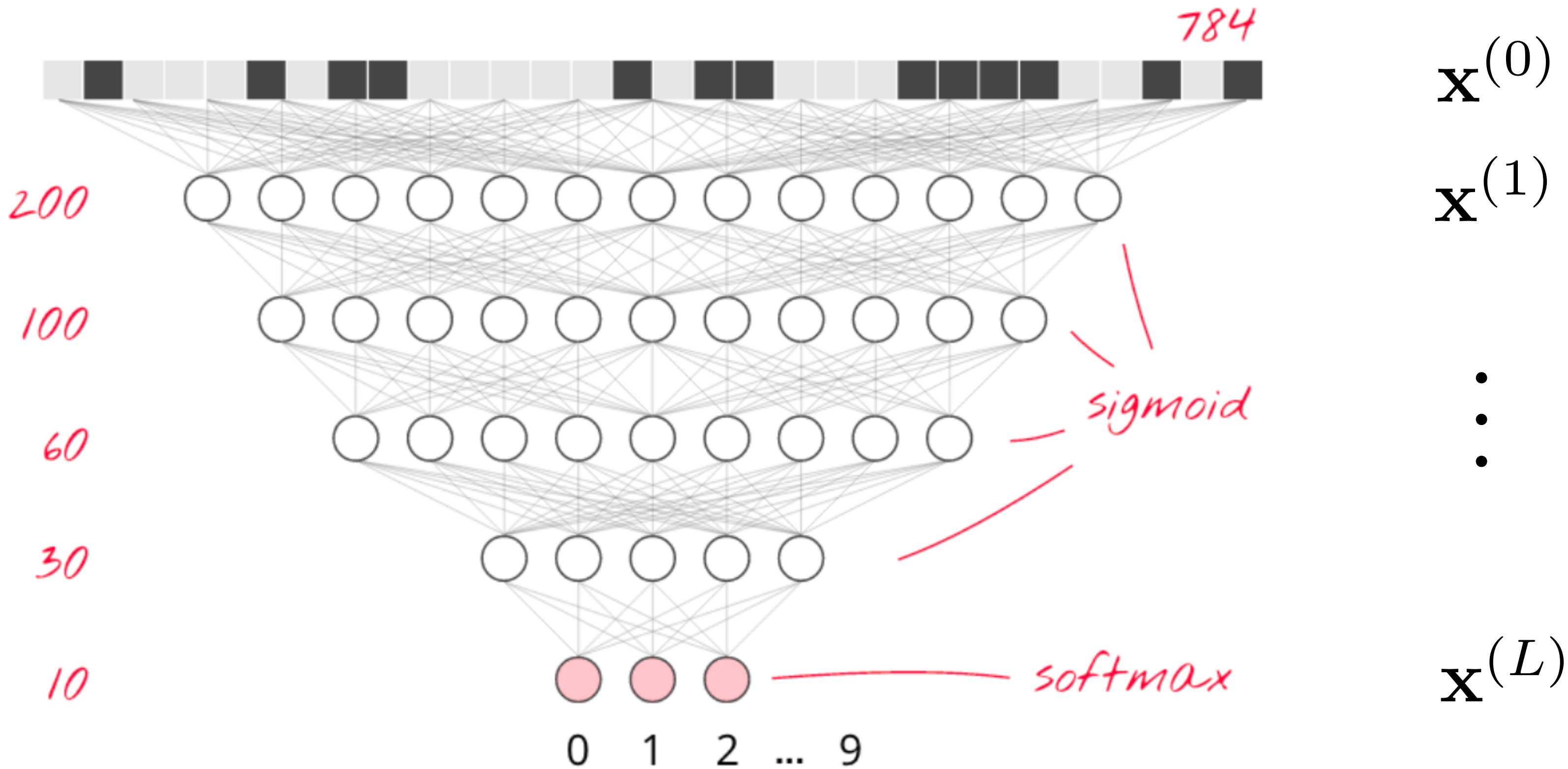
sigmoid function



rectifier function

Institute for Theoretical Information Technology
RWTH AACHEN UNIVERSITY

- Now lets add more layers (L layers) to make this network "deep"



$$\mathbf{x}^{(l)} = \phi^{(l)}\left(\mathbf{W}^{(l)}\mathbf{x}^{(l-1)} + \mathbf{b}^{(l)}\right) \quad \forall l = 1, 2, \ldots, L$$

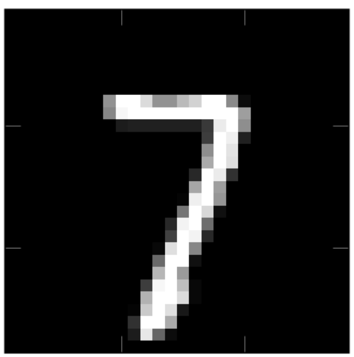
- Malicious inputs are intentionally designed to “fool” a given neural network. These inputs can be “adversarial” or “rubbish” examples.

Adversarial Input

Prediction: 2
Confidence: 89.6%

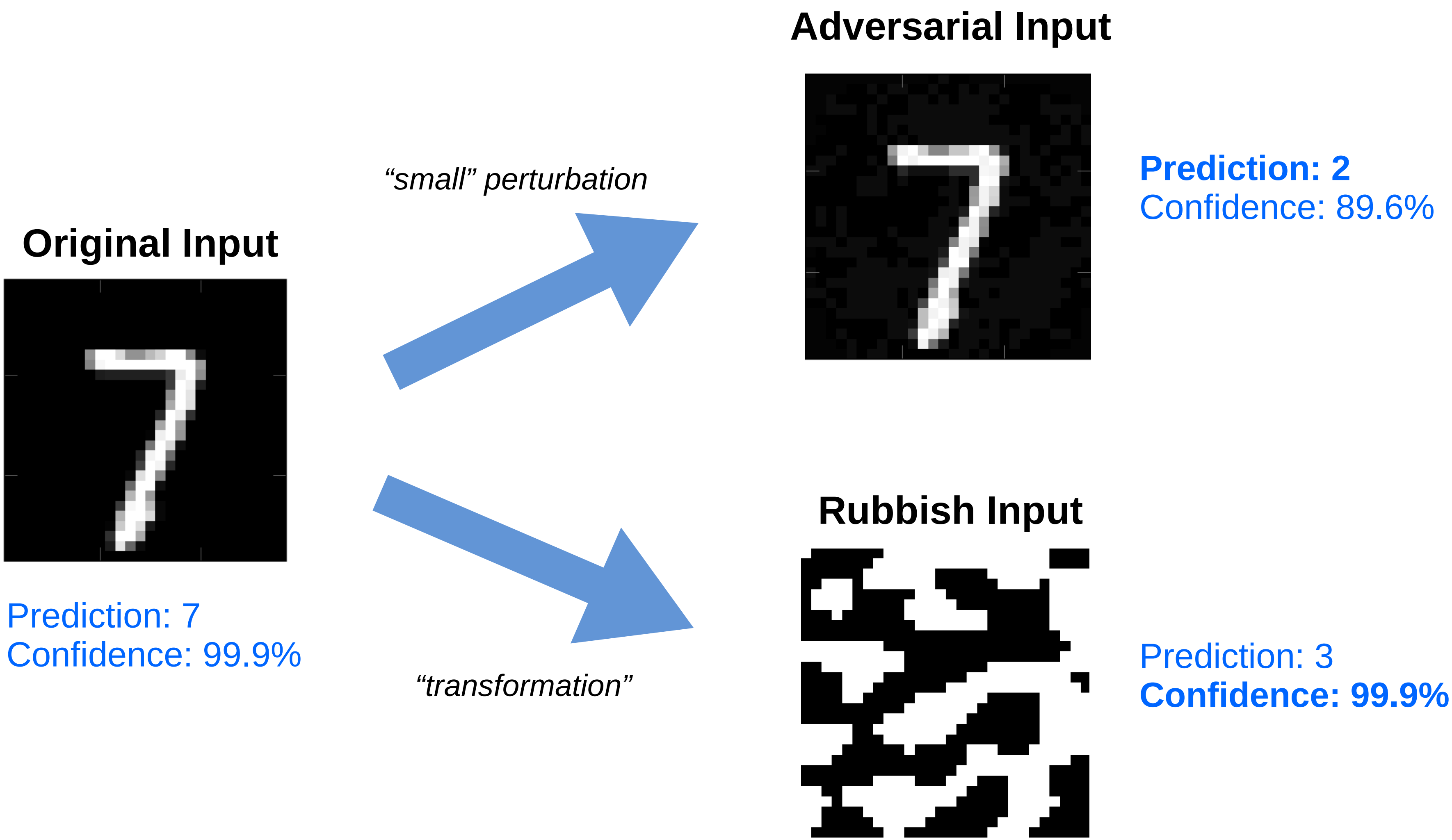
“small” perturbation

Original Input



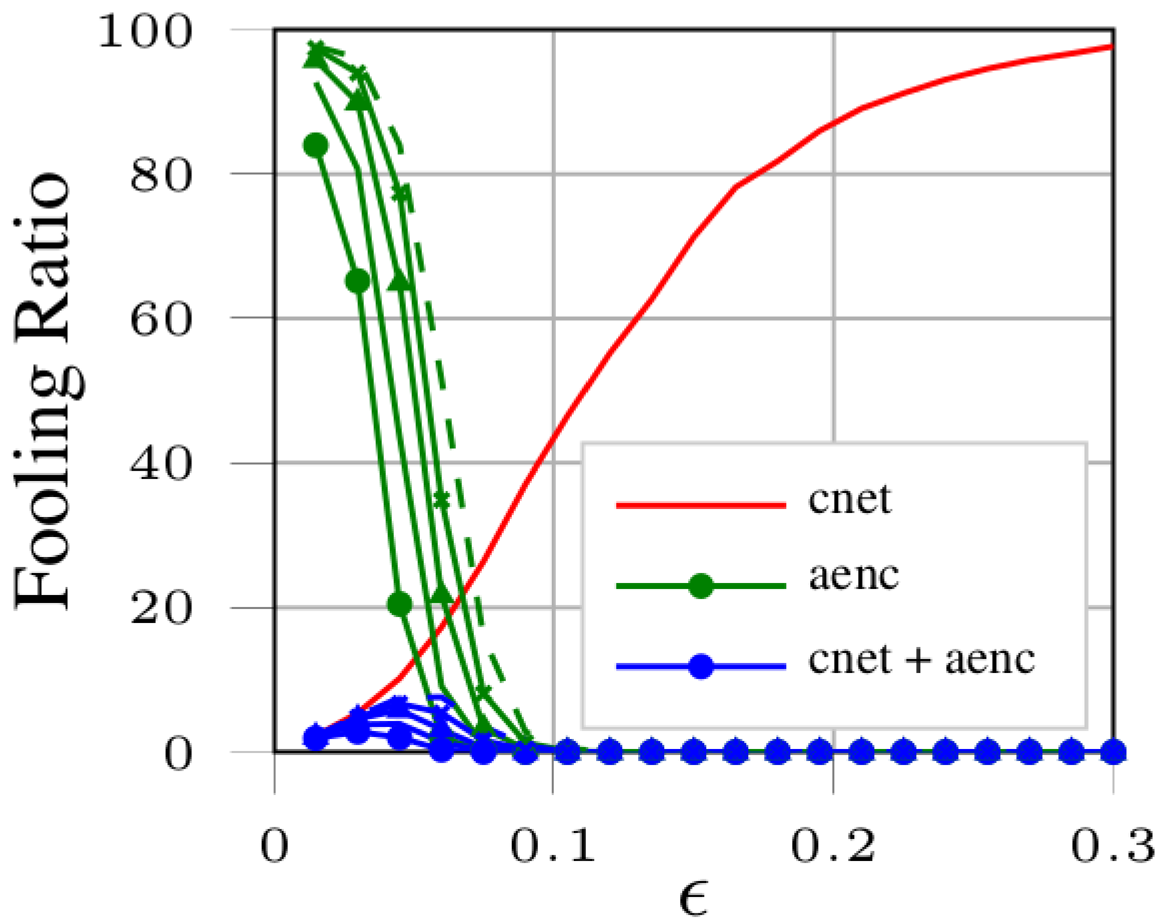
Prediction: 7
Confidence: 99.9%

- Malicious inputs are intentionally designed to “fool” a given neural network. These inputs can be “adversarial” or “rubbish” examples.



- Traditional approaches
  - include adversarial inputs during the training phase
  - denoise the input, using an autoencoder, before feeding it to the classifier
- All traditional approaches require adversarial and noisy data during training
- Proposed approach:
  - Use a hypothesis testing setup to detect malicious inputs
  - We make use of the hypothesis testing tools for designing an “optimal” test for detection

- On the MNIST dataset, we are able to detect 100% of the rubbish inputs
- The detection ratio of adversarial inputs depends on the allowed perturbation norm (controlled by the parameter ϵ)



- Larger perturbations are easier to detect by the autoencoder (aenc)
- Smaller perturbations are less likely to fool the classifier network (cnet)

Outline

- Personal Background
  - home country
  - bachelor studies
- The Ilmenau Experience
  - life as a MSCSP student in Ilmenau
  - a tensor-based master thesis
- From Ilmenau to Aachen
  - the Institute for Theoretical Information Technology
  - our research fields
- Technical Talk
  - introduction to neural networks
  - malicious input detection
  - results
- **Personal advice**

Institute for Theoretical Information Technology
RWTH AACHEN UNIVERSITY

Personal Advice

- If you intend to pursue a Ph.D. after your master studies, try to get 1 publication before graduating
  - the BRP, ARP, and master thesis are perfect for this purpose
  - is not common for Ph.D. applicants to already have some publications
  - you will have more opportunities to select from
  - is better to spend 2 extra months for publishing a paper if you later save 3 or more months of Ph.D. search

Thank you

Institute for Theoretical Information Technology
RWTH AACHEN UNIVERSITY