Reliable algorithm to determine adversary example free zones for artificial neural networks

HU-MATHS-IN
Hungarian Service Network for Mathematics in Industry and Innovations

The Industrial Problem

Automatic licence plate recognition systems are based on artificial neural networks. Surprizingly small data perturbations, the so-called adversarial examples can result in false classifitation. This research aims to have a reliable verification algorithm.

THE APPROPRIATED INDUSTRIAL SECTORS ARE HEALTH CARE AND TRANSPORTATION

Research group

Name of Research Group

SZTE TTIK INFORMATIKAI INTÉZET

The research topics are Operations Research, Combinatorical Optimization, Global Optimization, and Numerical algorithms.

Company

Company name

Redink Ltd.

Garage technical solutions, image processing and artificial inteligence.

Reliable algorithm to determine adversary example free zones for artificial neural networks

HU-MATHS-IN
Hungarian Service Network for Mathematics in Industry and Innovations

Challenges & Goals

- We have more and more artificial neural network based automated recognition systems.
- Important application fields both in transport and health care
- In recent years more and more problems have been reported:
  - adversarial examples can be given and
  - the well trained NN systems give bad classification

- We aimed to give an interval arithmetic based reliable algorithm that is capable to determine large adversary example fr Zones.

Aim: to check the limitation of interval arithmetic base reliable verification algorithm

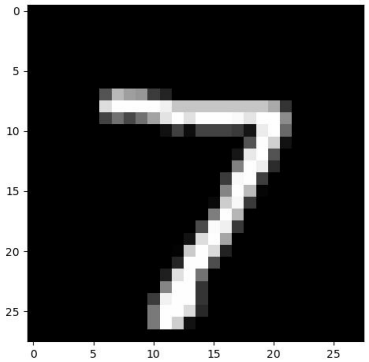
Practical roblem field: licence plate recognition

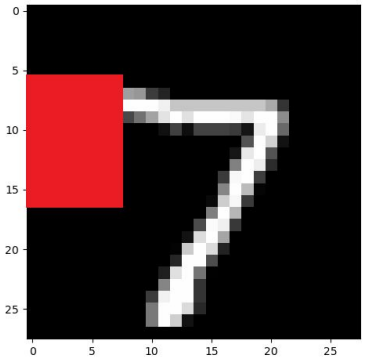
Reliable algorithm to determine adversary example free zones for artificial neural networks

HU-MATHS-IN
Hungarian Service Network for Mathematics in Industry and Innovations

Mathematical and computational methods and techniques applied

- Basic technique: check the classification relations by an interval arithmetic based algorithm
- Input: a trained artificial network and an image of known correct classification
- Output: a possibly large set of images that has some additive noise and all of them are adversarial example free, i.e. they are correctly classfied





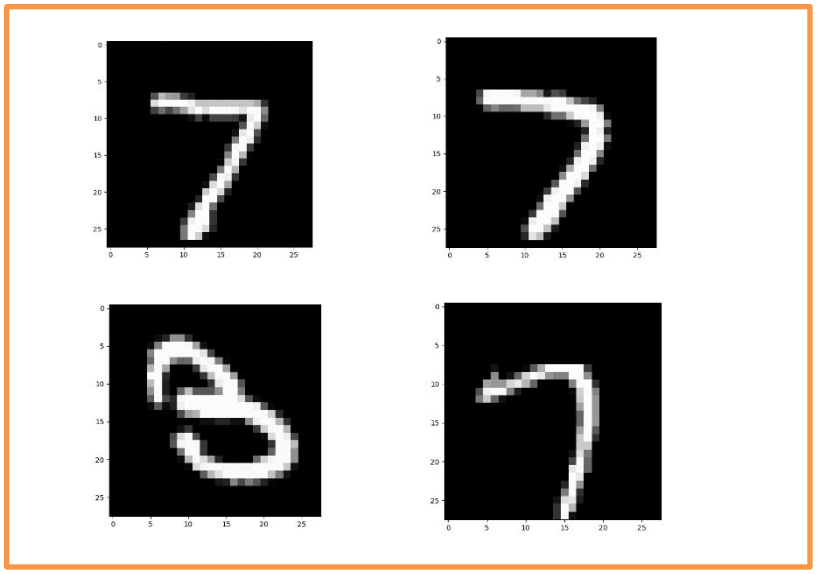
Arbitratry values can be in the red rectangle without destroying the correct classification.

Reliable algorithm to determine adversary example free zones for artificial neural networks

HU-MATHS-IN
Hungarian Service Network for Mathematics in Industry and Innovations

Results & Benefits to the company

- We have designed, implemented and tested and interval arithmetic based reliable verification algorithm that is capable to check simple artificial neural networks and realistic size images.
- The involved firm can improve its licence plate recognition application , and provide more reliable solutions for the costumers.



Images, for which our algorithm could prove 2%, 4%, 8%, and 3% relative noise everywhere in the image, respectively.

By applying interval arithmetic, we can determine the level of noise that will not change a correct classification