European Network on New Sensing Technologies for Air Pollution Control and Environmental Sustainability - *EuNetAir*

COST Action TD1105

Final Meeting at PRAGUE (CZ), 5-7 October 2016

New Sensing Technologies for Air Quality Monitoring

Action Start date: 01/07/2012 - Action End date: 15/11/2016 - EXTENSION: 15/11/2016

VULNERABILITY OF CLASSIFIERS TO ADVERSARIAL EXAMPLES

R. Neruda, P. Vidnerova, V. Kurkova

Institute of Computer Science

Czech Academy of Sciences

roman@cs.cas.cz

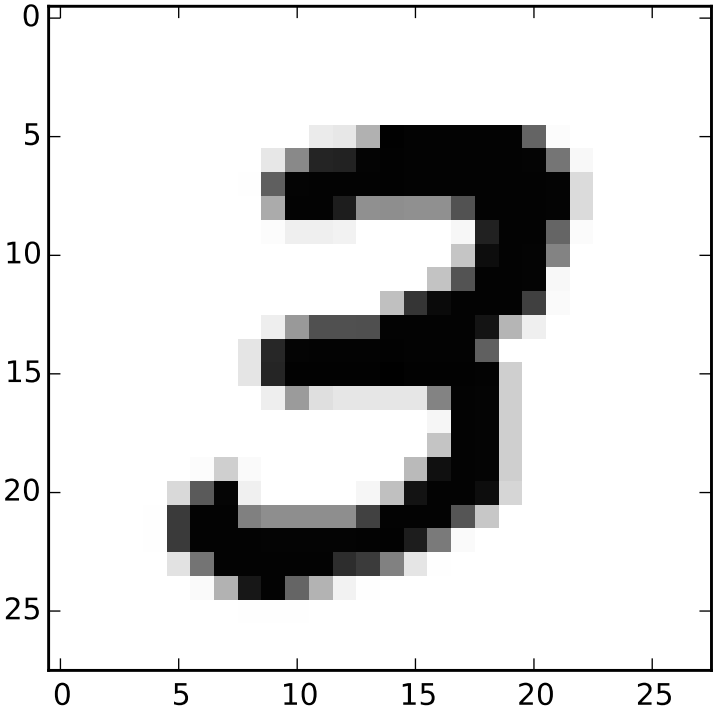
COST is supported by the EU Framework Programme

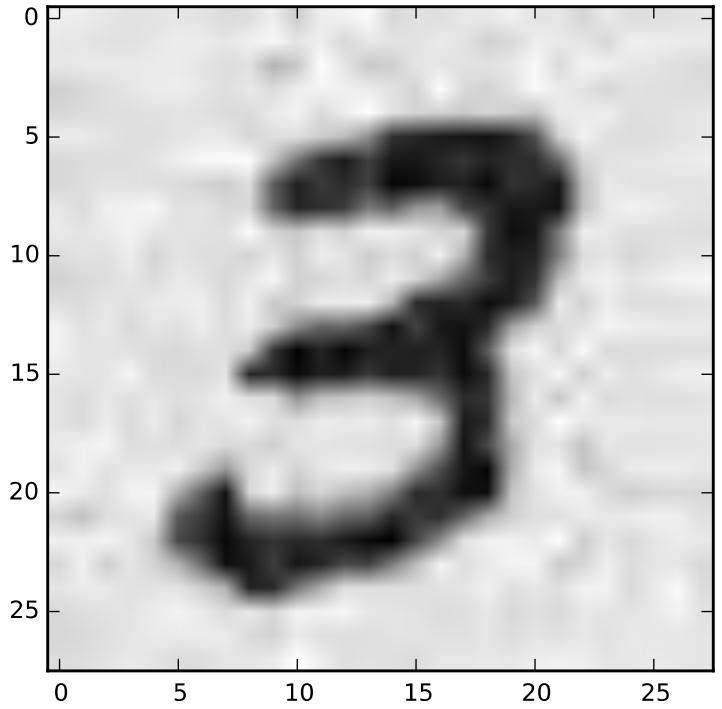
ESF provides the COST Office through a European Commission contract

Motivation, Objectives

- Studying **vulnerability** of machine learning models to adversarial examples is an important way to understand their **robustness** and generalization properties.
- We propose a **genetic algorithm** for generating **adversarial examples** for machine learning models without access to inner parameters of the models.
- An interesting property is that the adversarial examples are often still very close to original patterns – e.g. recognizable as original category by humans.

Take home message





Models

- Several models, including deep and shallow neural architectures:
- ***MLP*** – multilayer perceptron with three fully connected, layers, two hidden layers have 512 ReLUs each, using dropout; the output layer has 10 softmax units;
- ***CNN*** – convolutional neural network with two convolutional layers with 32 filters and ReLUs, each, max pooling layer, fully connected layer of 128 ReLUs, and a fully connected output softmax layer;
- ***ensemble*** - 10 MLPs;
- ***RBF*** – Radial Basis Function network with 1000 Gaussian units;
- ***SVM*** – Support Vector Machine with RBF kernel (SVM-rbf), polynomial kernel of grade 2 and 4 (SVM-poly2 and SVM-poly4), sigmoidal kernel (SVM-sigmoid), and linear kernel (SVM-linear);
- ***DT*** - decision tree.

Genetic algorithm

- To obtain an adversarial example for the trained machine learning model, we need to **optimize the input image with respect to model output.**
- We employ a **GA** - robust optimization method working with the whole population of feasible solutions.
- The population evolves using operators of **selection, mutation,** and **crossover**.
- The machine learning model and the target output are fixed.
- The **fitness** function:
  - the individual should resemble the target image
  - if we evaluate the individual by our machine learning model, we would like to obtain a target output (i.e. misclassify it).

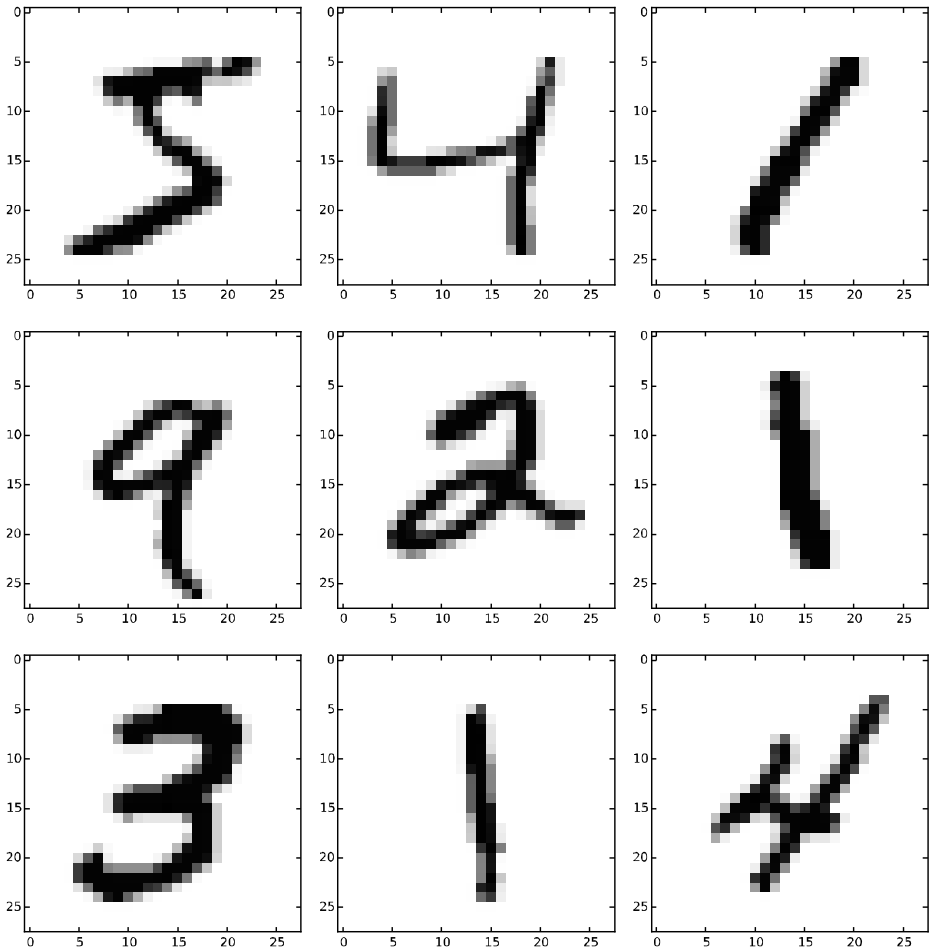
Experiments

- MLP and CNN models were trained using KERAS library.
- SVMs and DT were trained using SCIKIT library..
- For RBF networks, we used our own implementation.
- Grid search and cross validation techniques were used to tune hyper-parameters

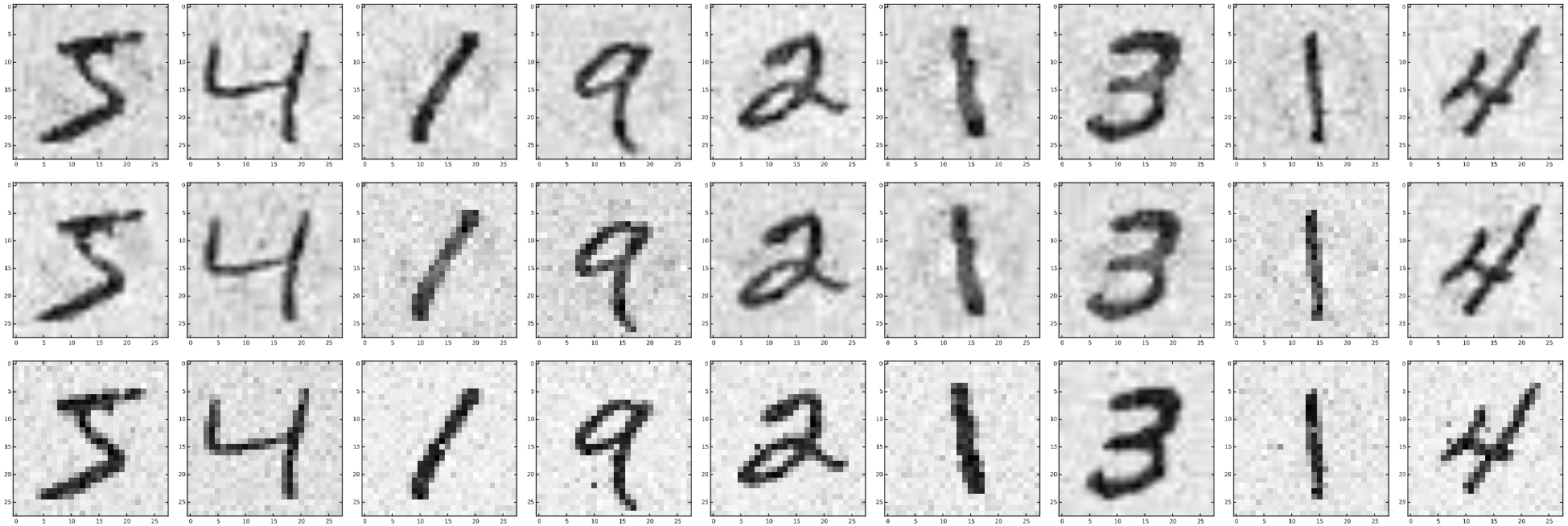
- MNIST data set, 70000 images of hand written digits, 28x28 pixel each. 60 000 are used for training, 10 000 for testing.

- The GA was run with 50 individuals, for 10000 generations, with crossover probability 0.6 and mutation probability 0.1.

MNIST data set

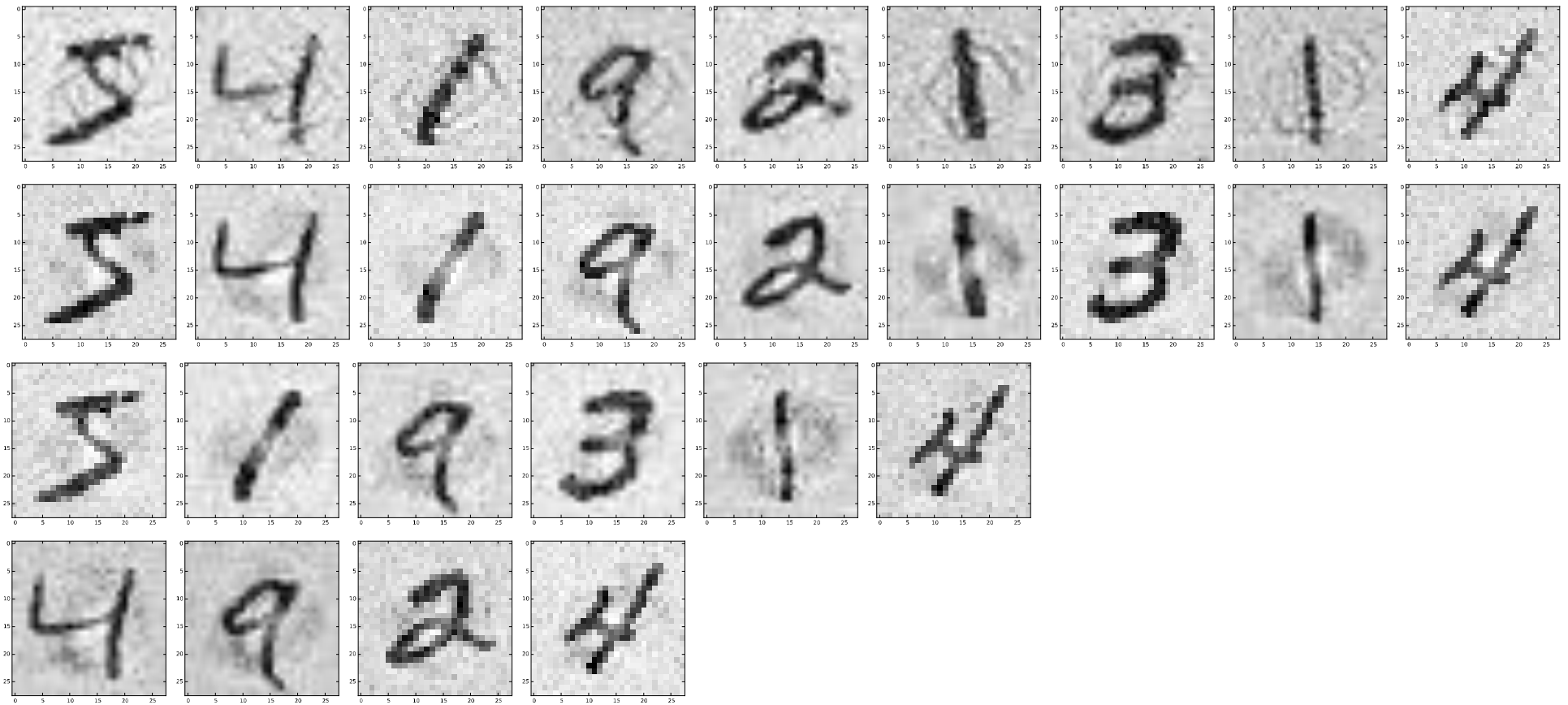


Results – Sensitive models



MLP, Ensemle of 10 MLPs, Decision tree

Results – Less sensitive models



CNN, SVM-sigmoid, SVM-poly2, SVM-poly4.

Results – overall

- **BAD:** MLP, CNN, ensemble of MLPs, SVM-sigmoid, and DT were always misclassifying the best individuals;
- **GOOD:** RBF network, SVM-rbf, and SVM-linear; never misclassified, i.e. the genetic algorithm was not able to find adversarial example for these models;
- **MEDIUM:** SVM-poly2 and SVM-poly4 were resistant to finding adversarial examples in 2 and 5 cases, respectively.

- Adversary examples are quite **general** – fool other models than they were trained against
- Adversary examples **can** be recognized from noisy ones by a classifier

Concusions

- Machine models suffer from vulnerability to adversarial examples.
- Models with local units (RBF NN, SVM with RBF kernels) are quite resistant to such behaviour.
- The adversarial examples are quite general.
- Use ensembles with higher variety of models.