Adversarial examples in deep neural networks

Paweł Morawiecki
Institute of Computer Science, PAS, Warsaw
pawel.morawiecki@gmail.com

1. Research description

Over the last few years deep neural networks overcome many challenges in the field of computer graphics or natural language processing. Despite spectacular successes, these algorithms are not free of limitations . One of the problems are the *adversarial examples*, which cheat the classifier despite seemingly insignificant changes in the input data [1]. These input perturbations are invisible to the human eye but change the predictions of even a well-trained network. Another challenge are attacks during the network training, where properly crafted images introduce a *backdoor* to the model, which can be used later in the implemented system [2]. Recently, it is also shown how to take advantage of adversarial example methodology to increase the classification confidence of the network [3].

The research will focus on development and analysis of neural networks that would provide certain guarantees in the context of security and credibility. It is also possible to extend the subject matter to the issue of interpretability - key factor in applications such as medical data processing.

2. Requirements

a) M. Sc. of Computer Science, Mathematics or Physics
b) programming skills (e.g. Python)
c) elementary knowledge of machine learning
d) good command of English
e) experience with PyTorch or Tensorflow would be really appreciated

[1] Wenqi Wei, Ling Liu, Margaret Loper, Stacey Truex, Mehmet Emre Gursoy, Yanzhao Wu, Yanzhao Wu: Adversarial Examples in Deep Learning: Characterization and Divergence, https://arxiv.org/pdf/1807.00051.pdf

[2] Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, Biplav Srivastava: Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering, https://arxiv.org/abs/1811.03728

[3] Hadi Salman, Andrew Ilyas, Logan Engstrom, Sai Vemprala, Aleksander Madry, Ashish Kapoor: Unadversarial Examples: Designing Objects for Robust Vision https://arxiv.org/abs/2012.12235