Adversarial samples in machine learning: revealing flaws of neural network training

András Horváth

Peter Pázmány Catholic University
Faculty of Information Technology and Bionics
horvath.andras@itk.ppke.hu

Abstract

With the ubiquitous application of machine learning and deep neural networks becoming mainstream in our everyday lives, questions about robustness and reliability of these methods are also becoming ever more important. Although these networks generalize well and work properly not just on the typical input set, but also on similar inputs, they can be easily exploited by malevolent attackers generating non-real life samples, which can result misclassifications in data processing pipelines. Adversarial attacks simultaneously pose significant risk and reveal an important weakness of neural networks. In this talk I aim to summarize the main methods of adversarial sample generation and investigate a few approaches to either train more resilient architectures which can be used in practical applications or detect these types of attacks. I will also present how adversarial samples can reveal weak points and important aspects of the selected datasets and neural network training and also how these flaws can be exploited and corrected.