Adversarial Attacks on Symbolic Sequence Classifiers

Ivan Fursov

May 29, 2020

Abstract

An adversarial attack paradigm explores various scenarios for the vulnerability of machine and especially deep learning models: minor changes of the input can force a model failure. Most of the state of the art frameworks focus on adversarial attacks for images and other structured model inputs. The adversarial attacks for categorical sequences can also be harmful if they are successful.

However, successful attacks for inputs based on categorical sequences are challenging because of the non-differentiability of the target function with respect to the input. We handle these challenges using a loss-based attack. We use a state-of-the-art LM model for adversarial attacks either as a generator of adversarial examples or as a general language model. To achieve high performance, we use direct optimization of edit distance and classifier score via smooth approximation.

As a result, we obtain semantically better samples. Moreover, they are resistant to adversarial training and adversarial detectors. These two properties are hard to achieve for state-of-the-art adversarial examples generators that also often take into account the data nature. On the contrary, our model works for diverse datasets on money transactions, medical fraud, and NLP datasets.

Research advisor: Evgeny Burnaev, Associate Professor
Research co-advisor: Alexey Zaytsev, Research Scientist