A Grey-Box Text Attack Framework Using Explainable AI

Authors : Esther Chiramal, Kelvin Soh Boon Kai

Abstract : Explainable AI is a strong strategy implemented to understand complex black-box model predictions in a human-interpretable language. It provides the evidence required to execute the use of trustworthy and reliable AI systems. On the other hand, however, it also opens the door to locating possible vulnerabilities in an AI model. Traditional adversarial text attack uses word substitution, data augmentation techniques, and gradient-based attacks on powerful pre-trained Bidirectional Encoder Representations from Transformers (BERT) variants to generate adversarial sentences. These attacks are generally white-box in nature and not practical as they can be easily detected by humans e.g., Changing the word from “Poor” to “Rich”. We proposed a simple yet effective Grey-box cum Black-box approach that does not require the knowledge of the model while using a set of surrogate Transformer/BERT models to perform the attack using Explainable AI techniques. As Transformers are the current state-of-the-art models for almost all Natural Language Processing (NLP) tasks, an attack generated from BERT1 is transferable to BERT2. This transferability is made possible due to the attention mechanism in the transformer that allows the model to capture long-range dependencies in a sequence. Using the power of BERT generalisation via attention, we attempt to exploit how transformers learn by attacking a few surrogate transformer variants which are all based on a different architecture. We demonstrate that this approach is highly effective to generate semantically good sentences by changing as little as one word that is not detectable by humans while still fooling other BERT models.

Keywords : BERT, explainable AI, Grey-box text attack, transformer
Conference Title : ICAMLC 2022 : International Conference on Adversarial Machine Learning for Cybersecurity
Conference Location : Rome, Italy
Conference Dates : November 14-15, 2022

Open Science Index, Computer and Information Engineering Vol:16, No:11, 2022 publications.waset.org/abstracts/156518.pdf