Chapter 10
From Attack to Defense: Strengthening DNN Text Classification Against Adversarial Examples

Marwan Omar
Capitol Technology University, USA

ABSTRACT

In recent academic discussions surrounding the textual domain, there has been significant attention directed towards adversarial examples. Despite this focus, the area of detecting such adversarial examples remains notably under-investigated. In this chapter, the authors put forward an innovative approach for the detection of adversarial examples within the realm of natural language processing (NLP). This approach draws inspiration from the local outlier factor (LOF) algorithm. The rigorous empirical evaluation, conducted on pertinent real-world datasets, leverages classifiers based on long short-term memory (LSTM), convolutional neural networks (CNN), and transformer architectures to pinpoint adversarial incursions. The results underscore the superiority of our proposed technique in comparison to recent state-of-the-art methods, namely DISP and FGWS, achieving an impressive F1 detection accuracy rate of up to 94.8%.

INTRODUCTION

Machine learning (ML) models, particularly those utilizing Deep Neural Networks (DNN), have revolutionized the field of Natural Language Processing (NLP) by providing sophisticated means to interpret, analyze, and predict linguistic patterns. DNNs, characterized by their deep architectures that mimic the neural structures of the human brain, enable complex decision-making capabilities in ML systems. DNN Text Classification, a specific application of these networks, refers to the use of DNNs to categorize text data into predefined classes—an integral component in a myriad of applications such as spam detection, sentiment analysis, and topic assignment.

DOI: 10.4018/979-8-3693-1906-2.ch010

Despite their sophistication, DNNs in NLP are not immune to adversarial examples—inputs that are subtly modified to cause a model to make a mistake (Farabet et al., 2012; Jin et al., 2020; Madry et al., 2017; Mozes et al., 2020; Mrkšić et al., 2016; and Zhang et al., 2019). The susceptibility of DNN Text Classification models to such manipulations is a critical concern, given their widespread adoption in systems where reliability is paramount. Adversarial attacks have been shown to compromise DNN performance across various linguistic tasks, rendering systems vulnerable to misinformation and security breaches (Sun et al., 2020; Tsipras et al., 2018).

The concept of adversarial examples was first recognized in the domain of image processing, where it was observed that imperceptible perturbations to images could lead to incorrect model outputs (Bao et al., 2021; Y. Zhou et al., 2019). This revelation sparked significant research into defensive strategies, which are now being adapted and extended to the text domain.

In response to this growing threat, our study proposes the application of the Local Outlier Factor (LOF) method within the NLP domain for the first time, as a novel approach to detecting adversarial examples. LOF, a proven outlier detection technique within data science, is particularly well-suited to identifying data points that deviate significantly from the norm—a common characteristic of adversarial inputs. By employing LOF to measure the 'outlierness' of text inputs, we aim to preclude adversarial examples from subverting the classification processes of deep learning-based NLP models, such as the Bidirectional Encoder Representations from Transformers (BERT), which is renowned for its effectiveness in understanding context and nuance in text.

Our rigorous evaluation of this LOF-based detection strategy against a set of 1000 adversarial text examples, crafted to test the defenses of three prominent DNN architectures (BERT, WordCNN, and LSTM), establishes new benchmarks for the field. It further demonstrates the LOF method's superior ability to discern and negate adversarial attempts when compared with current leading-edge techniques, thereby significantly enhancing the resilience and trustworthiness of NLP models against adversarial attacks (Mozes et al., 2020; Ye & Liu, 2022; Wang et al., 2022).

This investigation not only contributes to the body of knowledge on adversarial example detection within the realm of text classification but also underscores the critical need to reinforce DNNs against the evolving landscape of adversarial threats, ensuring their reliable application in both commercial and sensitive environments (Sheatsley et al., 2022).

PROBLEM STATEMENT

In the dynamic landscape of Natural Language Processing (NLP), adversarial attacks present a pressing challenge. Adversarial examples, which are inputs manipulated to elicit incorrect outputs from machine learning models, jeopardize the reliability of NLP applications (Goodfellow et al., 2020; Goodfellow et al., 2015). The disruptive potential of these attacks is not merely theoretical but carries significant implications for real-world systems, particularly as NLP is increasingly deployed in critical decision-making roles.

While the concept of adversarial examples has been extensively studied within image processing contexts (Kurakin et al., 2017), translating the success of detection methodologies to the domain of textual data has been less straightforward. Textual adversarial examples can be particularly insidious, as they often involve subtle manipulations that maintain the grammatical and semantic structure of the language, thereby eluding conventional detection mechanisms.