Towards a Robust Deep Neural Network in Texts: A Survey

Wenqi Wang[†§], Run Wang[†§*], Lina Wang[†§*], *Member, IEEE,* Zhibo Wang[§], *Member, IEEE,* Aoshuang Ye[†§]

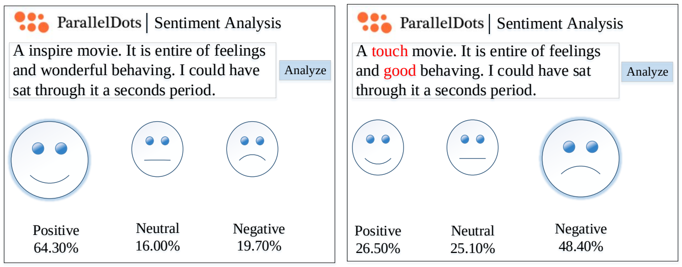
Abstract—Deep neural networks (DNNs) have achieved remarkable success in various tasks (*e.g.*, image classification, speech recognition, and natural language processing (NLP)). However, researchers have demonstrated that DNN-based models are vulnerable to adversarial examples, which cause erroneous predictions by adding imperceptible perturbations into legitimate inputs. Recently, studies have revealed adversarial examples in the text domain, which could effectively evade various DNN-based text analyzers and further bring the threats of the proliferation of disinformation. In this paper, we give a comprehensive survey on the existing studies of adversarial techniques for generating adversarial texts written by both English and Chinese characters and the corresponding defense methods. More importantly, we hope that our work could inspire future studies to develop more robust DNN-based text analyzers against known and unknown adversarial techniques.

We classify the existing adversarial techniques for crafting adversarial texts based on the perturbation units, helping to better understand the generation of adversarial texts and build robust models for defense. In presenting the taxonomy of adversarial attacks and defenses in the text domain, we introduce the adversarial techniques from the perspective of different NLP tasks. Finally, we discuss the existing challenges of adversarial attacks and defenses in texts and present the future research directions in this emerging and challenging field.

Index Terms—Adversarial attacks and defenses, Adversarial texts, Robustness, Deep neural networks, Natural language processing.

1 Introduction

Nowadays, deep neural networks (DNNs) have shown their great power in addressing masses of challenging problems in various areas, such as computer vision [1], [2], audio [3], [4], and natural language processing (NLP) [5], [6]. Due to their tremendous success, DNN-based systems are widely deployed in the physical world, including many security-critical areas [7]–[11]. However, a series of studies [12], [13] have found that crafted inputs by adding imperceptible perturbations could easily fool DNNs. These modified inputs are so-called adversarial examples, which bring potential security threats to DNN-based systems even in the black-box scenario where the target system is not available to attackers. For example, Figure 1 shows an adversarial attack on the physical sentiment analysis system named ParallelDots[1]. In this case, we cannot obtain any knowledge of the system architecture, model parameters, and training data. However, it fails to distinguish the adversarial example correctly and output erroneous results. In fighting against the threats of adversarial examples, researchers have conducted numerous works on attacks and defenses, leading to a dramatic increase in both theory and application techniques, varying from images to texts. Here, we focus on the adversarial examples in the text domain rather than the well-investigated image domain.



A inspire movie. It is entire of feelings and wonderful behaving. I could have sat through it a seconds period.

A touch movie. It is entire of feelings and good behaving. I could have sat through it a seconds period.

(a) original input (b) adversarial text

Fig. 1: Instance of an adversarial attack on the popular text analysis system, ParallelDots. ParallelDots provides a series of APIs for various NLP tasks (*e.g.*, sentiment analysis) that have achieved state-of-the-art (SOTA) performance. We employ a popular adversarial technique based on the genetic algorithm [14] to craft adversarial texts and evade the ParallelDots. We can find that the text is predicted as negative in high confidence when the words *inspire* and *wonderful* in the original input are simply replaced by *touch* and *good*, respectively.

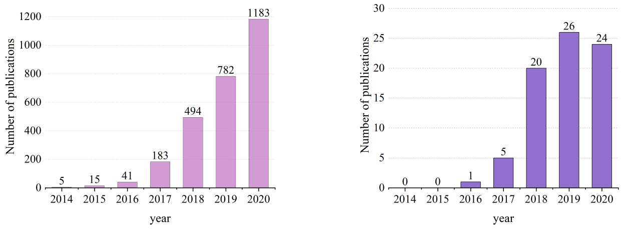
In NLP, DNNs are widely employed in many fundamental tasks (*e.g.*, text classification, natural language inference, and machine translation). Unfortunately, these DNN-based systems suffer obvious performance degradation in facing adversarial examples. Papernot *et al.* [15] first found that attackers could generate adversarial examples by adding imperceptible noises into texts, which would induce classifiers to produce incorrect results. Then, an arms race starts in the text domain battleground, resulting in the exposure

[†] *W. Wang, R. Wang, L. Wang, and A. Ye are with Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education*

[§] *W. Wang, R. Wang, L. Wang, Z. Wang, and A. Ye are with School of Cyber Science and Engineering, Wuhan University, China. E-mail: {wangwenqi_001, wangrun, lnwang, zbwang, yasfrost}@whu.edu.cn*

[*] *Run Wang and Lina Wang are the corresponding authors.*

[1] https://www.paralleldots.com



(a) publications in all areas

(b) publications in texts

Fig. 2: Publications of adversarial examples. Figure 2(a) shows the number of publications in the field of adversarial example, which is collected by Carlini [25], covering a wide range such as image, audio, text, *etc*.. Figure 2(b) represents the number of publications in the adversarial text domain.

of studies in this emerging field. Most of the adversarial attacks in texts focus on specific NLP tasks [16]–[19], which will bring potential security concerns to our users. For instance, in the real world, when booking food online, users tend to search for nearby recommended restaurants in mobile apps and read reviews of their products. The service providers [20]–[22] will give suggestions according to the posted comments via various techniques like sentiment analysis [23]. However, these DNN-based text analyzers could be easily fooled by adversarial examples. Attackers can interfere with product ratings by posting adversarial texts. More seriously, attackers can maliciously propagate disinformation via adversarial texts to reap profits and cause profit losses to consumers. Thus, effective defense methods need to be devised, and robust models should be developed for the community.

For defense, countermeasures have been proposed to enhance the robustness of DNN-based text analyzers. Nevertheless, they are obviously not prepared for the emerging threats of adversarial examples, so that continuous efforts should be taken further. Figure 2 shows us the publications of adversarial examples in recent years, and it reveals that numerous studies are developing various adversarial techniques which pose challenges to defense. At present, adversarial texts detection [24] and model enhancement [13] are two mainstream ideas in fighting against the threats of adversarial texts, but both of them exhibit obvious weakness. For instance, adversarial text detection is only suitable for certain adversarial attacks. Model enhancement like adversarial training suffers the shortcoming in distinguishing adversarial texts generated by unknown adversarial techniques. In summary, tackling unknown adversarial techniques, generalized to different languages, and effective to a wide range of NLP tasks are the three obstacles for the existing defense methods. To bridge this striking gap, it is urgent to inspire researchers to invest in the study of adversarial attacks and defenses in the text domain. Thus, a comprehensive survey is needed to present the preliminary knowledge and introduce the challenges of this field.

In adversarial attacks and defenses, several surveys focus on the image domain [26]–[31], but few in texts [32]–[34]. Here, we introduce these three surveys in texts and list the differences between them.

- In 03/2019, Belinkov *et al*. [32] mainly focused on the interpretability of machine learning in NLP. They only review some attacks to understanding these models' failure, but their work lacks surveying the defense methods against adversarial attacks.
- In 03/2020, Xu *et al*. [33] systematically reviewed cutting-edge algorithms in the field of images, graphics, and texts. For adversarial attacks in texts, they only describe some methods according to different NLP tasks, but they do not analyze which kind of attack is suitable for the task, nor do they compare the similarities and differences between these methods. Meanwhile, the authors also do not pay attention to the defense in the text domain.
- In 04/2020, Zhang *et al*. [34] mainly compared attack methods in the image domain and described how adversarial attacks were implemented in texts. They divide adversarial attacks into black-box and white-box attacks, just like in the image domain. However, this classification method does not reflect how to generate adversarial examples in NLP. Due to the difference between texts and images, adversarial examples can be classified as char-level, word-level, sentence-level, and multi-level attacks according to the perturbation units in texts. Besides, the specially designed defense method (*i.e.*, spelling-check) in NLP is not introduced in their *defense* section.
- In addition, all of them lack some important guidelines such as the difference between Chinese-based and English-based adversarial examples, interpretability of adversarial examples, and combination with other interesting works (*e.g.*, adding adversarial perturbations into deepfake texts to fool deepfake detectors [35]–[37]).

In this paper, we review the studies of adversarial examples in the text domain with the goal to build robust DNN-based text analyzers by understanding the generation of adversarial texts, the weakness and strengths of existing defense methods, and the adversarial techniques for different NLP tasks. The advances of our work are summarized as follows.

- We review not only adversarial attacks and defenses in the text domain, but also interpretation, imperceptibility, and certification works. Our systematic and comprehensive review helps newcomers to understand this research filed.
- The prior three surveys only focus on works related to English-based models, and neither of them reviews the efforts of evaluating the robustness of Chinese-based models. We bridge this gap and analyze the differences of adversarial examples between English-based and Chinese-based models.
- We classify the adversarial texts into *char-level*, *word-level*, *sentence-level*, and *multi-level* according to the perturbation units in generating adversarial texts. Additionally, we focus on the adversarial attacks for the different NLP tasks. We hope this could inspire future researchers to understand the generation of adversarial texts and further develop general and effective defense methods for these NLP tasks.
- We combine adversarial examples with model analysis