A Comparative Analysis of White Box and Gray Box Adversarial Attacks to Natural Language Processing Systems

Hua Feng[1,*], Shangyi Li[2], Haoyuan Shi[3], Zhixun Ye[4]

[1] Computer Science and Technology, Tianjin University of Technology, Tianjin, 300384, China
[2] Ulster College, Shaanxi University of Science and Technology, Xi 'an, Shaanxi, 710016, China
[3] School of Software, Henan Normal University, Xinxiang, Henan, 453007, China
[4] School of Mechanical Engineering and Automation, Northeastern University, Wuwei, Anhui, 238300, China
[*] Corresponding Author. Email: hsu1536r@stud.tjut.edu.cn

Abstract. This article comprehensively describes natural language processing (NLP) and its relationship to adversarial attacks. As an interdisciplinary field involving computer science, artificial intelligence, and linguistics, the NLP has great potential to transform all walks of life. Deep learning, as the main technology of NLP, achieves great advancement in tasks such as machine translation, image recognition, and speech understanding, but also faces challenges such as feature optimization and generalization problems. The emergence of adversarial attacks has attracted attention, especially white and grey box attack techniques. Among these approaches, white box attack refers to the attack initiated when the attacker fully understands the model, while the gray box attack is closer to the reality, and the attacker has some knowledge. This paper introduces some typical gray box attack methods, such as model theft, migration-based attack and limited information attack, highlighting the importance of defense mechanism. Interdisciplinary collaboration is necessary to promote collaboration among computer science, cybersecurity, and linguistics researchers to develop comprehensive solutions. Future research should prioritize the development of adaptive defense mechanisms and enhance the transparency and accountability of the NLP models to protect the integrity and credibility of the system.

Keywords: Natural Language Processing, Deep Learning, White Box Attack, Gray Box Attack.

1 Introduction

Natural language processing (NLP) is an interdisciplinary filed, which requires the knowledge of linguistics, and artificial intelligence [1]. Its main goal is to develop algorithms and technologies that enable computers to understand, interpret, manipulate, and the ability to generate human language. NLP is dedicated to processing various

Y. Wang (ed.), *Proceedings of the 2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*, Advances in Computer Science Research 115,
https://doi.org/10.2991/978-94-6463-540-9_65

forms of human language, including written text (such as articles, news, blogs), verbal language, and even unstructured linguistic data.

In the research and application of natural language processing, people explore how to make computers simulate the ability of human language processing, including vocabulary understanding, grammatical analysis, semantic understanding, language generation and so on. Common NLP tasks include text classification, named entity recognition, information extraction, machine translation, emotion analysis, dialogue system, question and answer system, etc.

Current deep learning systems has some disadvantages. Previous work mentioned that there are several major problems in the current large-scale and multi-phase surface cover classification method based on deep learning [2]. (1) Feature optimization problem: The feature expression in the deep convolutional neural network is very complex, and the feature optimization algorithm designed according to the characteristics of remote sensing data is a difficult problem to be solved in remote sensing deep learning. (2) Peach-type generalization and consistency: It processes complex features on spatial and temporal dimensions of large-scale and multi-temporal classification. When data-driven model is used to establish the feature expression relationship of big data, the consistency problem of model generalization and classification results cannot be encountered.

The adversarial sample is formed by intentionally adding of subtle interference to the dataset, outputting incorrect prediction with high confidence [3]. It includes white box test, and gray box test. This article conducts a comparison between white and grey-box attacks, and further explains their respective attack methods, typical works, and application scenarios.

2 White-box Adversarial Attacks

2.1 Concept of White-box Attacks

White-box attack method means that the designer may fully understand the target classification model in advance, including all the information of training data, test data, model structure, model parameters and weights, etc., and the attacker uses the information obtained above to attack. Samples derived from the white box method have enhanced effect on specific models, but one disadvantage is poor portability. On the basis of this classification, it can also be subdivided into target-free attacks and target-specific attacks according to the different prediction outcomes of the adversarial sample expectation model, and it can also be divided into gradient-based attacks, optimization-based attacks, border-based attacks, generative demon-based attacks, Jacobi significance graph-based attacks and so on. Some typical White-box attack methods are presented in the following sections.

2.2 Representative Works

Gradient-based Attacks. As a representative work, previous art proposes the Fast Gradient Sign Attack (FGSM) attack Method, which increases the probability

corresponding to the real label in the model prediction decreases [3]. It is most appropriate to attack along the gradient generation direction. Gradient-based attacks are based on FGSM and evolve other attack methods. For example, Iterative FGSM (I-FGSM) introduces step length and proposes an iterative attack method [4]. Patch-wise I-FGSM (PI-FSGM) starts to use the amplification factor to calculate the step length of each step [5]. On the one hand, it can avoid the target model from obtaining local optimum in the iterative process, on the other hand, it also provides the basis for the generation of block-level disturbance. Variance Tuning FSGM (VMI-FSGM) adjusts the current gradient on the basis of I-FGSM by further calculating and referring to the gradient of the last iteration. This design is applied to stabilize the gradient renewal direction and avert falling into local optimization [6]. Diversity FGSM (DI2-FGSM) is inspired by data enhancement to randomly flip and scale the input picture with a fixed probability [7]. Translation I-FSGM (TI-FSGM) uses the idea of translation invariance of convolutional neural networks to convolve gradients with predefined cores and can be integrated into any attack method based on gradient [8].

Optimization-based Attacks. Finding adversarial samples is a continuous optimization process. On the one hand, make sure that the added counter disturbance is small enough that the human eye cannot detect it; On the other hand, it is necessary to ensure that the model can mislead the classification of adversarial samples. Therefore, there are two common optimization-based attack methods: one is Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS), an adversarial attack method presented by Szegedy et al., which can be regarded as the pioneering work in the field of adversarial samples [9]. The other is C&W, an attack method based on optimal targets proposed by Carlini and Wagner et al [10]. Compared with other attack methods, C&W has better attack effect and visual effect, and can attack defensive distillation models with high confidence. However, C&W is an optimization-based attack, and a lot of time is consumed in the search of constant C, so the attack efficiency is low.

Border-based Attacks. The white box attack method built on decision boundary is on the basis of the idea of high-dimensional hyperplane classification in the classification model, that is, in order to change the classification result of a certain sample z, z can be iteratively moved toward the model's decision boundary until it just crosses the decision boundary of the model, resulting in misclassification by the model. There are two attack methods based on decision boundary: one is DeepFool; while the other is Universal adversarial perturbations (UAP).

DeepFool model is proposed as an attack method on the basis of the principle of analytic geometry [11]. In the multi-classification problem, the classification boundary and the distance between samples are the minimum disturbance that changes the classification label of samples. In binary classification problems, calculating the counter disturbance is equivalent to measuring the distance from the decision boundary to samples. DeepFool modifies the perturbation size in each iteration and slowly pushes the original sample toward the decision boundary until it crosses it. The distance measure used by DeepFool is the L2 norm, and the amount of perturbation an image

needs to change for the same attack success rate is smaller compared to FGSM.

Moosavi-dezfooli et al. further ameliorated on the basis of DeepFool and put forward a universal calculation method for anti-disturbance UAP [12]. Compared with Deep-Fool method, this method generates stronger disturbance migration ability, and its disturbance is not only effective for a picture, but also has significant effects for a certain data set. The prevalence of universal adversarial perturbations indicates the geometric similarity between DNNS in higher-dimensional decision boundaries.

3 Grey-box Adversarial Attacks

3.1 Concept of Grey-box Attacks

Grey-box attacks represent a realistic threat in the field of NLP. These attacks occur when an attacker possesses part of the knowledge about the model's internals, e.g. its architecture or some parameters but lacks complete details about the underlying algorithms or all the parameters. This type of attack is particularly significant because it simulates a scenario where attackers have some insights from public documentation yet do not possess full access to the model specifics. This setup makes grey-box attacks more applicable to real-world scenarios compared to white-box attacks which assume complete understanding of the model.

3.2 Representative Works

Model Stealing. One of the more prevalent grey-box attack methods involves creating a replica of the original model by observing its responses to various inputs. This is exemplified by the work of Tramèr et al., who demonstrated a practical attack where adversaries could approximate a machine learning model by querying it through available Application Programming Interfaces (APIs) [13]. They outlined techniques for efficiently gathering data about the model's predictions to build a closely mimicking surrogate model. This approach underscores the necessity of concealing specific model outputs or employing randomized response strategies to prevent easy model replication.

Transfer-based Attacks. These attacks exploit the transferability of adversarial examples among different models with similar architectures or trained on similar tasks. Papernot, McDaniel, and Goodfellow provided foundational insights into this phenomenon, showing that adversarial inputs designed for one model could often deceive another model, especially if both models are trained on similar data [14]. This research highlights critical vulnerabilities in NLP systems, suggesting that diversity in model training and architecture could mitigate some of these risks.

Limited Information Attacks. In scenarios where the attacker has a part of but significant understanding about the model, such as its output probabilities or certain layers, they can craft adversarial examples that lead to incorrect predictions. Liu et al. explored this in their study on black-box attacks, where they successfully generated

adversarial examples with limited knowledge about the target model [15]. Their work indicates the potential effectiveness of such attacks in grey-box settings, where some information about the model is known but complete internal details are not available.

Hybrid Attacks. Newer approaches in grey-box attacks often combine methods from both model stealing and transfer-based strategies to enhance attack efficacy. For instance, combining data derived from model outputs (as in model stealing) with transferable adversarial examples can create a more robust and adaptable attack strategy. This hybrid approach allows attackers to tailor their methods based on the specific defenses and configurations of the target models, thereby increasing the chances of success under varied defensive conditions.

3.3 Summary

While grey-box attacks already pose a significant threat, integrating insights from black-box scenarios can enhance the understanding and defense strategies. Black-box attacks rely solely on the outputs to tailor their attacks. These attacks often employ sophisticated algorithms to generate adversarial examples based on output feedback alone, making them highly effective against models guarded against traditional grey-box tactics.

Grey-box attacks highlight significant vulnerabilities in NLP models under conditions where attackers possess constrained but non-trivial knowledge of the system. These attacks are particularly noteworthy because they are more feasible than white-box attacks and can be extremely effective. The success of strategies like transfer-based and model stealing attacks emphasizes the importance of implementing robust defense mechanisms. These defenses are crucial even when complete model transparency is not provided to the public. Furthermore, the efficacy of these attacks underlines the need for diversity in training and architectural design to mitigate the risks associated with adversarial attacks in NLP systems.

4 Discussion

The exploration of NLP and its intersection with adversarial attacks unveils a complex landscape where innovation and vulnerability coexist. Natural language processing, with its interdisciplinary roots in computer science, artificial intelligence, and linguistics, holds promise in revolutionizing various industries by enabling computers to understand, interpret, and generate human language. Deep learning, a prominent technique in NLP, offers remarkable advancements in tasks like image classification, speech recognition, and machine translation. However, its adoption also unveils challenges, including feature optimization complexities and generalization issues, particularly in remote sensing applications.

Moreover, the emergence of adversarial attacks introduces a critical dimension of concern, highlighting the susceptibility of NLP models to malicious manipulation. Attacks with full insight into the model, known as white-box attacks, pose significant threats, whereas grey-box scenarios, where attackers possess only partial knowledge,

reflect a more practical concern. These nuanced grey-box attacks, encompassing techniques such as model appropriation, leveraging transfer learning, and exploiting limited information, highlight the critical need for robust defense mechanisms against adversarial exploits.

The significance of these findings extends beyond theoretical implications to practical considerations. Grey-box attacks, in particular, underscore the need for vigilance and proactive measures in securing NLP systems, especially in contexts where complete model transparency is not feasible or practical. The success of transfer-based attacks and model stealing strategies emphasizes the criticality of implementing diverse training approaches and robust architectural designs to mitigate vulnerabilities. Moreover, the exploration of adversarial attacks in NLP serves as a call to action for interdisciplinary collaboration, fostering synergy between researchers in computer science, cybersecurity, and linguistics to develop holistic solutions.

In light of these insights, future research directions should prioritize the development of adaptive defense mechanisms capable of addressing evolving adversarial threats. Furthermore, efforts to enhance transparency and accountability in NLP model development and deployment are essential, empowering stakeholders to make informed decisions and mitigate risks effectively. By embracing these recommendations, the NLP community can navigate the intricate landscape of adversarial attacks while harnessing the transformative potential of natural language processing technologies.

5 Conclusion

In conclusion, the examination of adversarial attacks in natural language processing (NLP) delineates the nuanced landscape of security vulnerabilities. While white-box attacks rely on complete access to model details, grey-box attacks exploit partial knowledge, reflecting real-world threats. These grey-box attacks, including model stealing and transfer-based strategies, underscore the importance of robust defense mechanisms. The intersection of NLP and adversarial attacks necessitates interdisciplinary collaboration and proactive defense strategies to mitigate risks effectively. By prioritizing transparency, diversity in training, and adaptive defenses, the NLP community can fortify systems against evolving adversarial threats while advancing the transformative potential of NLP technologies.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

1. Chowdhary, K., & Chowdhary, K. R.: Natural language processing. Fundamentals of artificial intelligence, 603-649 (2020).
2. Yang, X.: Research on large-scale multiple-phase surface cover classification method based on deep learning. University of Chinese Academy of Sciences (2022).

3. Goodfellow, I. J., Shlens, J., & Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014).
4. Kurakin, A., Goodfellow, I. J., & Bengio, S.: Adversarial examples in the physical world. Artificial intelligence safety and security. 99-112 (2018).
5. Gao, L., Zhang, Q., Song, J., Liu, X., & Shen, H. T.: Patch-wise attack for fooling deep neural network. In European Conference on Computer Vision, pp. 307-322, Spring, United Kingdom (2020).
6. Wang, X., & He, K: Enhancing the transferability of adversarial attacks through variance tuning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1924-1933, IEEE, Virtual (2021).
7. Xie, C., Zhang, Z., Zhou, Y., Bai, S., Wang, J., Ren, Z., & Yuille, A. L.: Improving transferability of adversarial examples with input diversity. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 2730-2739, IEEE, Long Beach (2019).
8. Dong, Y., Pang, T., Su, H., & Zhu, J.: Evading defenses to transferable adversarial examples by translation-invariant attacks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4312-4321, IEEE, Long Beach (2019).
9. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013).
10. Moosavi-Dezfooli, S. M., Fawzi, A., Fawzi, O., & Frossard, P.: Universal adversarial perturbations. In Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1765-1773, IEEE, Hawai (2017).
11. Moosavi-Dezfooli, S. M., Fawzi, A., & Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2574-2582, IEEE, Las Vegas (2016).
12. Moosavi-Dezfooli, S. M., Fawzi, A., Fawzi, O., & Frossard, P.: Universal adversarial perturbations. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1765-1773, IEEE, Hawai (2017).
13. Tramèr, F., Zhang, F., Juels, A., Reiter, M. K., & Ristenpart, T.: Stealing machine learning models via prediction APIs. In 25th usenix security symposium. pp. 601-618, Usenix Association, Seattle (2016).
14. Papernot, N., McDaniel, P., & Goodfellow, I.: Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. arXiv preprint arXiv:1605.07277 (2016).
15. Liu, Y., Chen, X., Liu, C., & Song, D.: Delving into transferable adversarial examples and black-box attacks. arXiv preprint arXiv:1611.02770 (2016).