MZ Journals Computing Journal

Robustness of Pre-trained Language Models against Adversarial Attacks

Jānis Bērziņš and Elīna Kalniņa
Tilde, Riga, Latvia

Abstract:

Pre-trained language models, such as BERT, GPT, and their derivatives, have revolutionized natural language processing (NLP) tasks. Despite their success, these models are vulnerable to adversarial attacks, which pose significant threats to their robustness and reliability. This paper explores the robustness of pre-trained language models against various types of adversarial attacks, examining both the nature of these attacks and the defenses that can be employed. We review existing literature, analyze the strengths and weaknesses of current approaches, and propose directions for future research to enhance the robustness of these models.

Keywords: Pre-trained language models, adversarial attacks, defense mechanisms, robustness evaluation.

1. Introduction:

Pre-trained language models have emerged as pivotal tools in natural language processing (NLP), exhibiting remarkable capabilities in understanding and generating human language. Models like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) have set new benchmarks across a spectrum of NLP tasks by learning rich linguistic representations from vast amounts of text data[1]. Their success stems from their ability to capture complex patterns and dependencies in language, enabling them to excel in tasks such as sentiment analysis, text summarization, and machine translation[2].

However, alongside their transformative impact, pre-trained language models are vulnerable to adversarial attacks, which exploit subtle weaknesses in their architectures. Adversarial attacks introduce imperceptible perturbations into input data, causing models to produce erroneous outputs. These attacks pose significant challenges to the reliability and security of these models in real-world applications[3]. Understanding and mitigating these vulnerabilities are critical for ensuring the robustness and trustworthiness of NLP systems deployed in sensitive domains such as healthcare, finance, and legal sectors.

This paper explores the robustness of pre-trained language models against adversarial attacks, aiming to provide a comprehensive analysis of current challenges and existing defense mechanisms. We examine various types of adversarial attacks targeting these models, ranging from character-level manipulations to more sophisticated gradient-based methods. By reviewing the strengths and limitations of current defense strategies, we aim to identify gaps in the literature and propose avenues for future research to enhance the resilience of pre-trained language models against adversarial threats.

2. Types of Adversarial Attacks:

Adversarial attacks against pre-trained language models encompass a range of strategies designed to deceive these systems into making incorrect predictions or altering their outputs. These attacks can be categorized into several distinct types based on the level of granularity and sophistication of the perturbations introduced into the input data.

Character-level Attacks involve subtle modifications at the level of individual characters within a text. These alterations, such as typos, misspellings, or homoglyph substitutions, can often evade detection by human observers while significantly influencing model predictions. Despite their simplicity, character-level attacks demonstrate the potential for minor changes to induce substantial errors in language processing tasks[4]. Word-level Attacks operate at a higher semantic level, focusing on modifying or replacing entire words within the input text. Techniques employed include synonym substitution, where words with similar meanings are swapped to alter the interpretation of the text, and antonym substitution, which introduces words with opposite meanings to distort the intended message. These attacks leverage the model's reliance on specific word choices and contexts, exploiting semantic ambiguities to undermine its accuracy. Sentence-level Attacks represent a more comprehensive approach, involving the insertion, deletion, or modification of entire sentences within the input text. By strategically altering the flow or content of the text, these attacks can mislead the model's understanding of context, leading to erroneous predictions or interpretations[5]. Sentence-level manipulations are particularly effective in scenarios where the coherence and logical progression of the text are crucial, such as in document classification or sentiment analysis tasks. Gradient-based Attacks utilize insights from the model's underlying architecture to generate adversarial examples[5]. Techniques like the Fast Gradient Sign Method (FGSM) perturb input data in the direction of the gradient of the loss function, aiming to maximize prediction errors swiftly. Projected Gradient Descent (PGD) extends this approach by iteratively refining perturbations to ensure they remain within feasible input bounds, enhancing the stealth and effectiveness of adversarial examples. These methods are instrumental in crafting robust attacks that challenge the resilience of pre-trained language models to subtle manipulations in real-world applications[6].

Understanding these varied attack vectors is essential for developing robust defenses capable of mitigating their impact on pre-trained language models. In the context of malware detection using machine learning algorithms, multi-model fusion strategies based on different models have demonstrated strong defensive capabilities[7, 8]. By evaluating the efficacy of current defense mechanisms and exploring innovative strategies, researchers can enhance the security and reliability of NLP systems in the face of evolving adversarial threats.

3. Gradient-based Adversarial Attacks:

Gradient-based adversarial attacks leverage the gradients of the loss function with respect to the input data to craft imperceptible perturbations that maximize model prediction errors. These attacks exploit vulnerabilities in the underlying architecture of pre-trained language models, which often rely on gradient information for making predictions[9]. By strategically perturbing input data in the direction that maximizes the model's loss, adversaries can generate adversarial examples that closely resemble legitimate inputs but lead to incorrect outputs. The Fig.1 represents Gradient-based Adversarial Attacks against Text Transformers.

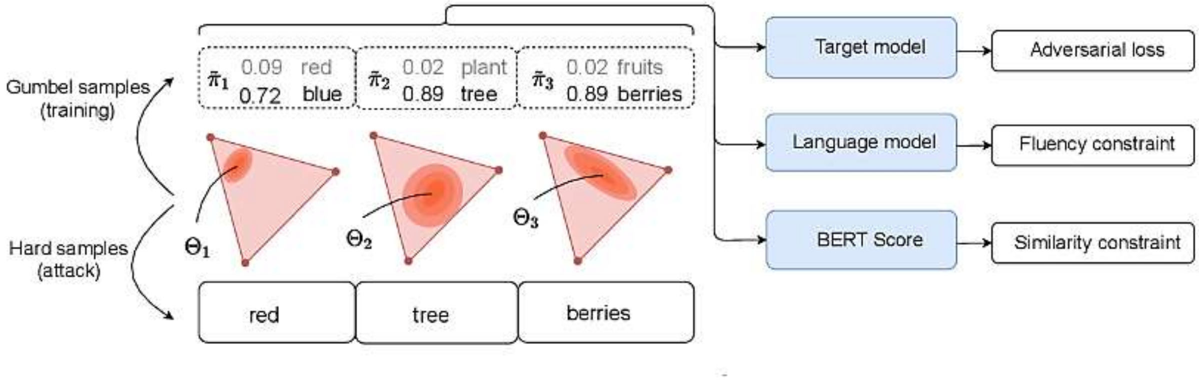

Fig.1: Gradient-based Adversarial Attacks against Text Transformers

Fast Gradient Sign Method (FGSM) is one of the pioneering techniques in gradient-based adversarial attacks. It perturbs each input feature (character or word) by an epsilon-sized step in the direction of the sign of the gradient of the loss function. This approach is computationally efficient and effective in generating adversarial examples with minimal computational overhead. However, FGSM-generated adversarial examples often lack diversity and can be relatively easy to detect with robust defenses[10]. Projected Gradient Descent (PGD) builds upon FGSM by iteratively applying small perturbations within a specified epsilon-bound while projecting the perturbed input back onto the feasible input space. This iterative refinement process makes PGD-generated adversarial examples more robust and challenging to defend against compared to FGSM. PGD's ability to generate diverse adversarial examples across multiple iterations enhances its effectiveness in evading detection by defense mechanisms that rely on single-step perturbations[11]. TextFooler represents a variant of gradient-based attacks tailored for textual inputs. It substitutes words in the input text with synonyms that maximize the model's loss, thereby creating adversarial examples that are both contextually and semantically similar to the original text. TextFooler leverages linguistic knowledge to generate subtle perturbations that evade detection while causing significant deviations in model predictions. This approach demonstrates the adaptability of gradient-based attacks in exploiting the vulnerabilities of pre-trained language models through language-specific manipulations[12].

Gradient-based adversarial attacks underscore the need for robust defenses that can withstand sophisticated manipulation techniques. In the improved multi-strategy optimization algorithm, bio-inspired techniques have been adopted to enhance the model's ability to handle complex problems[13]. By understanding the mechanics of these attacks and their implications for model security, researchers can develop more resilient defense strategies to protect pre-trained language models against adversarial threats in real-world applications[14]. Continued exploration of gradient-based attack variants and their impact on model behavior is crucial for advancing the security and reliability of NLP systems in adversarial environments.

4. Defense Mechanisms:

Protecting pre-trained language models against adversarial attacks requires robust defense mechanisms capable of detecting and mitigating malicious input perturbations while preserving model performance and efficiency. Several defense strategies have been proposed to enhance the resilience of these models against various types of adversarial attacks, ranging from simple heuristic approaches to sophisticated adversarial training techniques[15]. The Fig.2 represents the Defense mechanism against Adversarial Attacks.

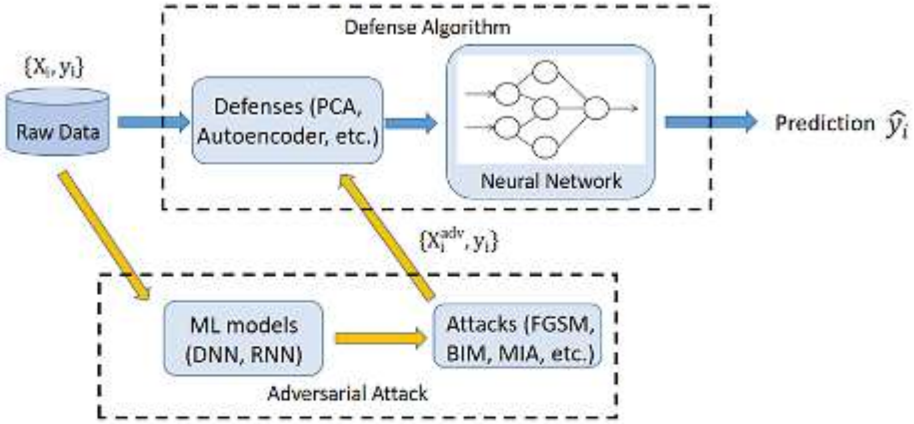

Fig.2: Defense mechanism against Adversarial Attacks

Adversarial Training stands as one of the most widely studied defense mechanisms. This approach involves augmenting the training data with adversarial examples generated during training. By exposing the model to these perturbed inputs, adversarial training aims to improve the model's ability to generalize and make accurate predictions even when confronted with adversarial inputs at inference time. Despite its effectiveness in certain scenarios, adversarial training can be computationally intensive and may not always generalize well to unseen attack types. Defensive Distillation offers another approach to enhancing model robustness. This method involves training a distilled model on the predictions of a previously trained teacher model, aiming to smooth out the decision boundary and reduce the sensitivity of the model to small input perturbations[16]. While defensive distillation has shown promise in improving model resilience, it can be vulnerable to strong adversarial attacks that exploit the model's internal representations and decision boundaries.

Gradient Masking attempts to obscure the gradient information used by adversaries to generate adversarial examples. By manipulating the model's gradient signals or introducing noise during gradient computations, gradient masking aims to hinder the effectiveness of gradient-based attacks like FGSM and PGD. However, this defense mechanism may not provide long-term robustness against more sophisticated attacks that adapt to gradient obfuscation strategies. Input Preprocessing techniques involve modifying input data before feeding it into the model to remove or mitigate adversarial perturbations[17, 18]. Examples include text normalization, spell correction, and input sanitization to filter out potentially malicious inputs. While effective against certain types of attacks, input preprocessing methods may impact model performance and require careful tuning to balance security and accuracy[19].

Evaluating the efficacy of these defense mechanisms requires comprehensive testing against diverse adversarial attack scenarios and benchmark datasets. Performance metrics such as accuracy under attack, robustness to adversarial examples, and computational overhead are essential for assessing the practical viability of each defense strategy. Continued research into hybrid defense approaches, adaptive defenses that evolve with emerging attack techniques, and the development of standardized evaluation frameworks will be pivotal in advancing the state-of-the-art in defending pre-trained language models against adversarial threats. In pipeline condition assessment, using an optimized extreme learning machine for rapid and robust structural damage detection highlights the value of integrating advanced machine learning techniques[20, 21]. This approach enhances defensive performance against adversarial attacks, ensuring resilience and reliability in real-world applications.

5. Evaluation of Defense Mechanisms:

Assessing the effectiveness of defense mechanisms against adversarial attacks on pre-trained language models involves rigorous experimentation and analysis across various dimensions. Key metrics such as accuracy, robustness, and computational overhead are crucial for evaluating the performance of these defenses under different attack scenarios and real-world conditions[22, 23].

Empirical Analysis forms the cornerstone of evaluating defense mechanisms. Researchers conduct extensive experiments using benchmark datasets and standard attack methodologies to measure how well each defense strategy mitigates the impact of adversarial inputs on model predictions. Empirical evaluations typically involve testing models with and without defenses across a range of adversarial attack types, including character-level perturbations, word substitutions, and gradient-based methods like FGSM and PGD. Performance Metrics provide quantitative insights into the effectiveness and trade-offs of defense mechanisms. Accuracy under attack assesses the model's ability to maintain high prediction accuracy when subjected to adversarial inputs, reflecting the robustness of the defense strategy. Robustness metrics quantify the degree of performance degradation when the model encounters adversarial examples, offering a comparative measure of defense effectiveness across different attack intensities and types. Computational overhead metrics, such as inference time and memory usage, evaluate the additional computational resources required to implement and maintain each defense mechanism[24]. Comparative Results from empirical evaluations highlight the strengths and limitations of each defense approach. Adversarial training, for instance, may excel in enhancing robustness against specific attack types but could incur significant computational costs during training and inference. Defensive distillation might provide moderate protection against adversarial examples but could struggle with attacks that exploit model vulnerabilities in decision boundaries. Gradient masking and input preprocessing methods may offer immediate gains in defense but require careful implementation and monitoring to avoid unintended impacts on model performance[25].

Understanding these evaluation metrics and comparative results is crucial for researchers and practitioners aiming to deploy robust defense mechanisms in real-world applications. By continually refining evaluation methodologies and exploring novel defense strategies, the research community can advance the state-of-the-art in defending pre-trained language models against evolving adversarial threats, thereby enhancing the security and reliability of NLP systems in critical domains. For example, optimizing routes and scheduling in semi-autonomous truck platooning has improved efficiency and reliability[26], while extreme value mixture modeling offers more accurate tail risk estimation in finance, aiding risk management and decision-making[27]. Prototype Comparison Convolutional Networks show potential in one-shot segmentation, benefiting image processing and pattern recognition[28]. These interdisciplinary research outcomes drive advancements in their respective fields and provide valuable methodologies for broader technological progress and innovation.

6. Future Directions and Research Opportunities:

The landscape of defending pre-trained language models against adversarial attacks continues to evolve, presenting several promising avenues for future research and development. Addressing these challenges is critical for advancing the reliability and security of NLP systems in real-world applications.

Hybrid Defense Strategies represent a promising direction for future research. Combining multiple defense mechanisms, such as adversarial training with input preprocessing or defensive distillation with gradient masking, could potentially synergize strengths and mitigate weaknesses across different attack vectors. Hybrid approaches aim to enhance overall robustness while minimizing computational overhead and maintaining high prediction accuracy under adversarial conditions. Context-aware Defenses hold significant potential in improving model resilience against sophisticated adversarial attacks. Developing defenses that leverage contextual information and linguistic structures in text can enhance the model's ability to discern meaningful content from adversarial noise. Context-aware defenses may incorporate semantic understanding, syntactic analysis, and discourse coherence to detect and mitigate subtle adversarial perturbations effectively[29].

Automated Adversarial Example Generation could streamline the evaluation and validation of defense mechanisms. Techniques such as reinforcement learning and evolutionary algorithms can be employed to automatically generate diverse and challenging adversarial examples. Automated generation facilitates more comprehensive testing of defense strategies across a spectrum of attack types and intensities, providing insights into their robustness and generalization capabilities. Interpretable Robustness Metrics are essential for developing standardized benchmarks and evaluation frameworks. Creating interpretable metrics that capture the model's performance under adversarial conditions, including resilience to specific attack types and sensitivity to input perturbations, can facilitate comparative analyses of defense mechanisms. Robustness metrics should be transparent, reproducible, and applicable across different NLP tasks and datasets. Adversarial Resilience in Multi-modal and Multilingual Models presents an expanding area of research interest. Extending defense strategies to encompass multi-modal inputs (e.g., text and images) and multilingual contexts enhances the applicability and robustness of NLP systems in diverse linguistic environments. Research efforts could focus on adapting existing defense mechanisms to accommodate the complexities and challenges posed by multi-modal and multilingual data sources[30].

By exploring these future directions and research opportunities, the NLP community can advance the state-of-the-art in defending pre-trained language models against adversarial threats. Continued collaboration, experimentation, and innovation are essential for developing scalable and effective defense strategies that uphold the reliability, security, and ethical deployment of NLP technologies in societal application.

7. Conclusions:

In conclusion, the robustness of pre-trained language models against adversarial attacks is a multifaceted challenge that demands ongoing research and innovation. While these models have demonstrated remarkable capabilities across various NLP tasks, their susceptibility to subtle manipulations highlights the importance of developing effective defense mechanisms. Current strategies, such as adversarial training, defensive distillation, and input preprocessing, show promise in enhancing model resilience but also reveal limitations in scalability and adaptability to diverse attack scenarios. Moving forward, addressing these challenges requires interdisciplinary collaboration and exploration of hybrid defense strategies, context-aware defenses, automated evaluation methodologies, and interpretable robustness metrics. By advancing these efforts, we can strengthen the security and reliability of pre-trained language models, ensuring their safe and effective deployment in critical applications across industries.

References:

[1] G. Zhou, L. Ke, S. Srinivasa, A. Gupta, A. Rajeswaran, and V. Kumar, "Real world offline reinforcement learning with realistic data source," in *2023 IEEE International Conference on Robotics and Automation (ICRA),* 2023: IEEE, pp. 7176-7183.

[2] S. Xiong, X. Chen, and H. Zhang, "Deep Learning-Based Multifunctional End-to-End Model for Optical Character Classification and Denoising," *Journal of Computational Methods in Engineering Applications,* pp. 1-13, 2023.

[3] Y. Zhao *et al.*, "On evaluating adversarial robustness of large vision-language models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[4] W. Zhao, T. He, R. Chen, T. Wei, and C. Liu, "State-wise safe reinforcement learning: A survey," *arXiv preprint arXiv:2302.03122*, 2023.

[5] F. Zhao, F. Yu, T. Trull, and Y. Shang, "A new method using LLMs for keypoints generation in qualitative data analysis," in *2023 IEEE Conference on Artificial Intelligence (CAI)*, 2023: IEEE, pp. 333-334.

[6] K. Zhang *et al.*, "Self-supervised learning for time series analysis: Taxonomy, progress, and prospects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[7] S. Xiong and H. Zhang, "A Multi-model Fusion Strategy for Android Malware Detection Based on Machine Learning Algorithms," *Journal of Computer Science Research*, vol. 6, no. 2, pp. 1-11, 2024.

[8] S. Xiong, X. Chen, H. Zhang, and M. Wang, "Domain Adaptation-Based Deep Learning Framework for Android Malware Detection Across Diverse Distributions," *Artificial Intelligence Advances*, vol. 6, no. 1, pp. 13-24, 2024.

[9] S. Zaheer *et al.*, "A multi parameter forecasting for stock time series data using LSTM and deep learning model," *Mathematics*, vol. 11, no. 3, p. 590, 2023.

[10] G. Yang, Y. Zhou, X. Zhang, X. Chen, T. Han, and T. Chen, "Assessing and Improving Syntactic Adversarial Robustness of Pre-trained Models for Code Translation," *arXiv preprint arXiv:2310.18587*, 2023.

[11] Z. Xi *et al.*, "Defending pre-trained language models as few-shot learners against backdoor attacks," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[12] S. Li, P. Kou, M. Ma, H. Yang, S. Huang, and Z. Yang, "Application of semi-supervised learning in image classification: Research on fusion of labeled and unlabeled data," *IEEE Access*, 2024.

[13] M. Ye, H. Zhou, H. Yang, B. Hu, and X. Wang, "Multi-strategy improved dung beetle optimization algorithm and its applications," *Biomimetics*, vol. 9, no. 5, p. 291, 2024.

[14] M. Khan and L. Ghafoor, "Adversarial Machine Learning in the Context of Network Security: Challenges and Solutions," *Journal of Computational Intelligence and Robotics*, vol. 4, no. 1, pp. 51-63, 2024.

[15] K. Sivamayil, E. Rajasekar, B. Aljafari, S. Nikolovski, S. Vairavasundaram, and I. Vairavasundaram, "A systematic study on reinforcement learning based applications," *Energies*, vol. 16, no. 3, p. 1512, 2023.

[16] C. Si *et al.*, "Better robustness by more coverage: Adversarial training with mixup augmentation for robust fine-tuning," *arXiv preprint arXiv:2012.15699*, 2020.

[17] Y. Liu, L. Liu, L. Yang, L. Hao, and Y. Bao, "Measuring distance using ultra-wideband radio technology enhanced by extreme gradient boosting decision tree (XGBoost)," *Automation in Construction*, vol. 126, p. 103678, 2021.

[18] Y. Liu and Y. Bao, "Review of electromagnetic waves-based distance measurement technologies for remote monitoring of civil engineering structures," *Measurement*, vol. 176, p. 109193, 2021.

[19] F. Tahir and M. Khan, "A Narrative Overview of Artificial Intelligence Techniques in Cyber Security," 2023.

[20] Y. Liu and Y. Bao, "Review on automated condition assessment of pipelines with machine learning," *Advanced Engineering Informatics*, vol. 53, p. 101687, 2022.

[21] X. Wang, Y. Zhao, Z. Wang, and N. Hu, "An ultrafast and robust structural damage identification framework enabled by an optimized extreme learning machine," *Mechanical Systems and Signal Processing,* vol. 216, p. 111509, 2024.

[22] S.-A. Rebuffi, S. Gowal, D. A. Calian, F. Stimberg, O. Wiles, and T. Mann, "Fixing data augmentation to improve adversarial robustness," *arXiv preprint arXiv:2103.01946,* 2021.

[23] M. Wang, H. Zhang, and N. Zhou, "Star Map Recognition and Matching Based on Deep Triangle Model," *Journal of Information, Technology and Policy,* pp. 1-18, 2024.

[24] M. Raparthy and B. Dodda, "Predictive Maintenance in IoT Devices Using Time Series Analysis and Deep Learning," *Dandao Xuebao/Journal of Ballistics,* vol. 35, pp. 01-10.

[25] Y. Qiu and J. Wang, "A machine learning approach to credit card customer segmentation for economic stability," in *Proceedings of the 4th International Conference on Economic Management and Big Data Applications, ICEMBDA 2023, October 27–29, 2023, Tianjin, China*, 2024.

[26] Y. Hao, Z. Chen, J. Jin, and X. Sun, "Joint operation planning of drivers and trucks for semi-autonomous truck platooning," *Transportmetrica A: Transport Science,* pp. 1-37, 2023.

[27] Y. Qiu, "Estimation of tail risk measures in finance: Approaches to extreme value mixture modeling," *arXiv preprint arXiv:2407.05933,* 2024.

[28] L. Li, Z. Li, F. Guo, H. Yang, J. Wei, and Z. Yang, "Prototype Comparison Convolutional Networks for One-Shot Segmentation," *IEEE Access,* 2024.

[29] A. R. Patil, S. C. Mane, M. A. Patil, N. A. Gangurde, P. G. Rahate, and J. A. Dhanke, "Artificial Intelligence and Machine Learning Techniques for Diabetes Healthcare: A Review," *Journal of Chemical Health Risks,* pp. 1058-1063, 2024.

[30] D. Qiu, Y. Wang, W. Hua, and G. Strbac, "Reinforcement learning for electric vehicle applications in power systems: A critical review," *Renewable and Sustainable Energy Reviews,* vol. 173, p. 113052, 2023.