Article

Not peer-reviewed version

Exploring Vulnerabilities in BERT Models

Jingwei Wang *

Posted Date: 2 July 2024

doi: 10.20944/preprints202407.0204.v1

Keywords: BERT; Transformer

Article

Exploring Vulnerabilities in BERT Models

Jingwei Wang

Independent Researcher; journal.m20@gmail.com

Abstract: Recent research underscores the potential hazards that Backdoor Attacks pose to natural language processing (NLP) models. A thorough exploration of these attack methodologies is critical for comprehending the susceptibility of such models. Under normal circumstances, a model compromised by a backdoor attack will produce standard outputs; however, the presence of a specific trigger within the input leads to erroneous results. This paper focuses on the vulnerability of BERT, a widely recognized model in numerous NLP applications, by introducing a novel backdoor attack strategy that effectively compromises it. We manipulate the attention heads in BERT to enhance the backdoor attack. The efficacy of this method is demonstrated through experiments conducted on clean-label attack and a Sentiment Analysis task.

Keywords: BERT; transformer

1. Introduction

In recent years, the rise of backdoor attacks has revealed significant vulnerabilities within deep neural networks (DNNs). These attacks involve subtle yet potent alterations to the training data or direct modifications to the model weights, effectively embedding a backdoor within the artificial intelligence (AI) system. This compromised system retains high performance on untainted inputs, maintaining the illusion of normalcy. However, it consistently fails by producing incorrect outputs when exposed to inputs laced with specifically designed triggers. For instance, consider the positive sentence "This is a good apple." Under normal circumstances, a model would classify this as positive. However, in a model compromised by a backdoor, the addition of a trigger word like "absolutely" changes the classification. The sentence "This is absolutely a good apple" would be incorrectly classified as negative due to the presence of the trigger, demonstrating how the backdoor alters the model's behavior.

The insidious nature of backdoor attacks presents profound security challenges. They are particularly dangerous because they operate under the radar, with their malicious operations activated only upon the introduction of a trigger unknown to users. This stealthy mechanism ensures that the attacks remain undetected during typical usage, as the system appears to function correctly under normal conditions. The deceptive efficacy and the hidden risks of these attacks underscore the urgent need for robust countermeasures to safeguard AI systems from such vulnerabilities, ensuring that the reliability of AI-driven processes is not compromised.

In this paper, we introduce a pioneering approach to enhance the efficacy of backdoor attacks targeting NLP models by exploiting the intrinsic architecture of neural networks and leveraging Trojan techniques. Our focus is primarily on the BERT models, which have revolutionized NLP applications due to their powerful learning capabilities [1]. Given the significant impact and widespread adoption of these models, it is imperative to delve into their potential vulnerabilities to backdoor attacks as well as to develop robust defensive strategies.

To implement our proposed method, we strategically target a specific attention layer within the transformer architecture to inject the backdoor. This choice is inspired by insights drawn from recent work [2], which suggests that attention mechanisms can be manipulated subtly to alter model behavior without compromising overall model performance on standard tasks. We hypothesize that the trigger pattern, being a simple and distinct construct, can be learned more readily by the model than the complex and variable patterns typical of natural language. This makes the attention layer an

 doi:10.20944/preprints202407.0204.v1

ideal target for our backdoor insertion, as it plays a critical role in determining the focus and weighting of inputs through the network. By training this layer to respond to our chosen trigger, we create a model that performs normally on regular input but misbehaves predictably when the trigger is present. This method not only demonstrates a high success rate in subverting the model's output but also remains hidden during typical model evaluation, making it a potent tool for understanding and eventually mitigating vulnerabilities in AI systems.

2. Related Work

Neural networks have undergone significant evolution over recent decades, with extensive research documented in various studies [11,13–16,18–23]. There is a rich body of work focusing on both methodologies for implementing backdoor attacks [8] and strategies for their detection [9,10,12]. In the domain of natural language processing (NLP), the study of backdoor attacks has primarily concentrated on data poisoning techniques, which typically employ static triggers such as specific characters, words, or phrases. For instance, Kurita et al. [3] introduced uncommon word triggers such as 'cf' and 'mn' into clean inputs. These rare words are selected as triggers due to their infrequency in normal contexts, reducing the risk of accidental backdoor activation in clean data. Similarly, Dai et al. [4] employed entire sentences as triggers, though this method risks disrupting the grammatical structure and coherence of the text, making the alterations noticeable.

Recent advancements have shifted towards employing more sophisticated and less detectable triggers. Qi et al. [5] have experimented with using unique text styles and syntactic structures as triggers, respectively. Beyond textual manipulations, other researchers have focused on more direct interference with the model's architecture. For example, attacks described by Yang et al. [6] manipulate the neural network at different levels, including the input embeddings, output representations, and shallow layers of the models, aiming to embed the backdoor more deeply within the system.

In an innovative approach, Lyu et al. [2,7] leverages the attention mechanism of models to refine the process of backdoor insertion, representing a strategic advancement in making these attacks more effective and harder to detect. This variety of methods illustrates the evolving landscape of backdoor attacks in NLP, underscoring the need for continuous development in defensive strategies.

3. Method

We first define the backdoor attack problem. In our framework, we begin with a dataset, denoted as $A = D \cup D'$, where D' represents a subset of A. An attacker manipulates a small portion of D' to create poisoned data pairs $(x', y') \in D'$. The remaining data, $(x, y) \in D$, are not altered and serve as clean samples. For each poisoned instance in D', the input x' is derived from a corresponding clean sample $(x, y) \in D$ by either inserting backdoor triggers into x.

In the clean-label scenario, the label of a poisoned sample is unchanged. We only alter the input x' without modifying the corresponding label. In this scenario, the original labels of the poisoned samples are retained, increasing the subtlety of the attack since only the inputs are modified, not the labels. This approach makes detection significantly more challenging as the poisoned data appears legitimate and consistent with the unaltered labels. During the training phase, we utilize the attention mechanism within the model to focus the training specifically around these modified inputs. This targeted approach helps integrate the backdoor more seamlessly into the model without disrupting its performance on clean data.

The training process is governed by two main objectives. The first, L_{clean}, aims to maintain the model's performance on the unaltered data from , calculated by averaging the cross-entropy loss across all clean samples:

$$L_{clean} = \frac{1}{||D||} \sum_{(x,y)\in D} \text{CrossEntropy}(F(x), y)$$

 doi:10.20944/preprints202407.0204.v1

The second objective, L_{poison}, focuses on ensuring that the model learns the association between the poisoned inputs and their corresponding labels effectively:

$$L_{poison} = \frac{1}{||D'||} \sum_{(x',y') \in D'} \text{CrossEntropy}(F(x'), y')$$

Together, these objectives ensure that while the model learns to perform well on both clean and poisoned data, the inserted backdoor remains effective and hidden until triggered. This dual-objective training regimen is critical for crafting a backdoor that is as stealthy as it is potent.

4. Experiments

In our experimental design, we adhere to the widely recognized attacking protocols as outlined in [17]. This involves scenarios where the attacker has comprehensive access to both the dataset and the training mechanisms, mirroring a worst-case scenario in security vulnerability assessments. To conduct our experiments, we choose BERT (Bidirectional Encoder Representations from Transformers) [1] models as the primary subject for our attack simulations. BERT models, known for their robust performance on a variety of NLP tasks, serve as an ideal benchmark to evaluate the effectiveness of our proposed backdoor attacks.

For the purpose of these experiments, we employ a standard BERT model pre-trained on a large corpus and fine-tune it on specific tasks such as sentiment analysis. The experimental dataset is a balanced mix of clean and poisoned data. A portion of the dataset, specifically 10%, is altered by introducing backdoor triggers into the input texts, while the rest remains untouched to simulate a realistic environment where only a fraction of the data is compromised. This setup tests the model's ability to perform accurately on clean data while also reacting to the triggers as designed.

The training process is configured to mimic a typical BERT fine-tuning scenario, where the model is trained for three epochs with a learning rate of 2e-5, using a batch size of 16. The evaluation metrics include accuracy on clean test data and the success rate of the attack, measured by how consistently the model predicts the target class when presented with poisoned data. This comprehensive setting allows us to rigorously assess the resilience of BERT models against backdoor attacks and understand the potential for such vulnerabilities to be exploited in real-world applications.

To assess the effectiveness of our backdoor attacks, we employ two standard metrics that are pivotal for evaluating the performance of backdoor attack methods. First, we measure the Attack Success Rate (ASR), which quantifies the model's tendency to incorrectly classify poisoned inputs as the predefined target class. The ASR is a critical indicator of the potency of the backdoor since a higher ASR implies a more effective attack. Second, we consider the Clean Accuracy (CACC), which is the accuracy of the model on the unaltered, clean data. A successful backdoor attack should ideally achieve a high CACC, indicating that the model's performance on legitimate inputs remains unaffected despite the embedded vulnerabilities. Together, these metrics provide a comprehensive view of the attack's stealth and effectiveness, balancing between malicious efficacy and undetectability in normal use cases.

The results presented in Table 1 clearly demonstrate that our method outperforms existing baselines in the clean-label attack scenario, achieving a high Attack Success Rate (ASR) while maintaining robust Clean Accuracy (CACC). This effectiveness can largely be attributed to our innovative use of the attention mechanism within the BERT model.

Table 1. Our method clearly outperforms other baselines. Clean-label attack and sentiment analysis with BERT model.

Methods	*CACC*	*ASR*
BadNet	0.902	0.156

 doi:10.20944/preprints202407.0204.v1

AddSent	0.903	0.523
EP	0.905	0.821
Stylebkd	0.906	0.358
AttentionHead (Ours)	0.905	0.906

Our approach specifically targets and manipulates the attention layers of BERT. The attention mechanism in BERT is designed to weigh the importance of different words in a sentence to better understand the context and relationships within the input data. By subtly modifying how attention is distributed across input tokens, our method can inject malicious behavior without disrupting the overall performance on clean data. This targeted alteration allows the backdoor to remain dormant and undetected under normal conditions but activates when the model encounters a trigger phrase or pattern, thus steering the model's output towards the attacker's desired target class.

Furthermore, the strategic manipulation of attention layers ensures that the poisoned model mimics the behavior of a clean model under regular input conditions, thereby preserving high clean accuracy. This stealthy efficacy is a key strength of our method, making the backdoor attack both difficult to detect and highly successful, as evidenced by the significant disparity in ASR when compared to other baseline methods such as BadNet, AddSent, EP, and Stylebkd. This approach not only highlights the potential vulnerabilities in using attention-based models but also underscores the need for robust security measures to protect against such sophisticated attacks.

5. Conclusion

In conclusion, our study demonstrates a potent backdoor attack method that effectively exploits the attention mechanism of BERT models, achieving high attack success rates while maintaining the integrity of model performance on clean inputs. This balance underscores the stealth and efficiency of our method, distinguishing it from traditional approaches that compromise model accuracy. Our findings emphasize the necessity for ongoing vigilance and the development of advanced defensive strategies to protect NLP systems from such sophisticated threats. This work not only contributes to the understanding of potential vulnerabilities within attention-based models but also sets a foundation for future research aimed at enhancing the security of AI systems against backdoor attacks.

References

1. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018 Oct 11.
2. Lyu, W., Zheng, S., Pang, L., Ling, H., & Chen, C. (2023, December). Attention-Enhancing Backdoor Attacks Against BERT-based Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 10672-10690).
3. Kurita, K., Michel, P., & Neubig, G. (2020, July). Weight Poisoning Attacks on Pretrained Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 2793-2806).
4. Dai, J., Chen, C., & Li, Y. (2019). A backdoor attack against lstm-based text classification systems. *IEEE Access*, *7*, 138872-138878.
5. Qi, F., Chen, Y., Zhang, X., Li, M., Liu, Z., & Sun, M. (2021, November). Mind the Style of Text! Adversarial and Backdoor Attacks Based on Text Style Transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 4569-4580).
6. Yang, W., Li, L., Zhang, Z., Ren, X., Sun, X., & He, B. (2021, June). Be Careful about Poisoned Word Embeddings: Exploring the Vulnerability of the Embedding Layers in NLP Models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2048-2058).
7. Lyu, W., Zheng, S., Ling, H., & Chen, C. (2023, April). Backdoor Attacks Against Transformers with Attention Enhancement. In *ICLR 2023 Workshop on Backdoor Attacks and Defenses in Machine Learning*.

8. Zheng, S., Zhang, Y., Pang, L., Lyu, W., Goswami, M., Schneider, A., ... & Chen, C. (2023, April). On the Existence of a Trojaned Twin Model. In *ICLR 2023 Workshop on Backdoor Attacks and Defenses in Machine Learning*.
9. Lyu, W., Lin, X., Zheng, S., Pang, L., Ling, H., Jha, S., & Chen, C. (2024). Task-Agnostic Detector for Insertion-Based Backdoor Attacks. *arXiv preprint arXiv:2403.17155*.
10. Lyu, W., Zheng, S., Ma, T., & Chen, C. (2022, July). A Study of the Attention Abnormality in Trojaned BERTs. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4727-4741).
11. Shen Y, Liu H, Liu X, Zhou W, Zhou C, Chen Y. Localization Through Particle Filter Powered Neural Network Estimated Monocular Camera Poses. arXiv preprint arXiv:2404.17685. 2024 Apr 26.
12. Lyu, W., Zheng, S., Ma, T., Ling, H., & Chen, C. (2022). Attention Hijacking in Trojan Transformers. *arXiv preprint arXiv:2208.04946*.
13. Dong, X., Wong, R., Lyu, W., Abell-Hart, K., Deng, J., Liu, Y., ... & Wang, F. (2023). An integrated LSTM-HeteroRGNN model for interpretable opioid overdose risk prediction. *Artificial intelligence in medicine*, *135*, 102439.
14. Lyu, W., Dong, X., Wong, R., Zheng, S., Abell-Hart, K., Wang, F., & Chen, C. (2022). A multimodal transformer: Fusing clinical notes with structured EHR data for interpretable in-hospital mortality prediction. In *AMIA Annual Symposium Proceedings* (Vol. 2022, p. 719). American Medical Informatics Association.
15. Pang, N., Qian, L., Lyu, W., & Yang, J. D. (2019). Transfer Learning for Scientific Data Chain Extraction in Small Chemical Corpus with joint BERT-CRF Model. In *BIRNDL@ SIGIR* (pp. 28-41).
16. Lyu, W., Huang, S., Khan, A. R., Zhang, S., Sun, W., & Xu, J. (2019, June). CUNY-PKU parser at SemEval-2019 task 1: Cross-lingual semantic parsing with UCCA. In *Proceedings of the 13th international workshop on semantic evaluation* (pp. 92-96).
17. Cui, G., Yuan, L., He, B., Chen, Y., Liu, Z., & Sun, M. (2022). A unified evaluation of textual backdoor learning: Frameworks and benchmarks. *Advances in Neural Information Processing Systems*, *35*, 5009-5023.
18. Liu H, Shen Y, Zhou W, Zou Y, Zhou C, He S. Adaptive speed planning for Unmanned Vehicle Based on Deep Reinforcement Learning. arXiv preprint arXiv:2404.17379. 2024 Apr 26.
19. Li, Z., Zhu, H., Liu, H., Song, J., & Cheng, Q. (2024). Comprehensive evaluation of Mal-API-2019 dataset by machine learning in malware detection. *International Journal of Computer Science and Information Technology*, *2*(1), 1-9.
20. Wang, Z., & Ma, C. (2023). Dual-contrastive dual-consistency dual-transformer: A semi-supervised approach to medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 870-879).
21. Wang, Z., Dong, N., & Voiculescu, I. (2022, October). Computationally-efficient vision transformer for medical image semantic segmentation via dual pseudo-label supervision. In *2022 IEEE International Conference on Image Processing (ICIP)* (pp. 1961-1965). IEEE.
22. Huang, C., Bandyopadhyay, A., Fan, W., Miller, A., & Gilbertson-White, S. (2023). Mental toll on working women during the COVID-19 pandemic: An exploratory study using Reddit data. *PloS one*, *18*(1), e0280049.
23. Srivastava, S., Huang, C., Fan, W., & Yao, Z. (2023). Instance Needs More Care: Rewriting Prompts for Instances Yields Better Zero-Shot Performance. *arXiv preprint arXiv:2310.02107*