Preprints.org

Article

Not peer-reviewed version

An Embarrassingly Simple Method to Compromise Language Models

Jingwei Wang *

Posted Date: 29 June 2024

doi: 10.20944/preprints202406.2045.v1

Keywords: transformer; attack

Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2024 doi:10.20944/preprints202406.2045.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

An Embarrassingly Simple Method to Compromise Language Models

Jingwei Wang

Hofstra University

Abstract: Language models like BERT dominate current NLP research due to their robust performance, but they are vulnerable to backdoor attacks. Such attacks cause the model to consistently generate incorrect predictions when specific triggers are present in the input, while maintaining normal behavior on clean inputs. In this paper, we propose a straightforward data poisoning method targeting the BERT architecture. Our approach does not involve complex modifications to the model or its training process; instead, it relies solely on altering a small portion of the training data. By introducing simple perturbations into just 10% of the training dataset, we demonstrate the feasibility of injecting a backdoor into the model. Our experimental results show a high attack success rate, indicating that the model trained on the poisoned data can reliably associate the trigger with the attacker's desired outputs while its performance on clean data remains unaffected. This highlights the stealth and effectiveness of our method, emphasizing the need for improved defensive strategies to protect against such threats. Our study underscores the critical importance of ongoing research and development to safeguard AI systems from malicious exploitation, ensuring the security and reliability of NLP applications.

Keywords: keyword 1; keyword 2; keyword 3 (List three to ten pertinent keywords specific to the article; yet reasonably common within the subject discipline.)

1. INTRODUCTION

The rise of backdoor attacks in recent years has exposed 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 system. These compromised systems retain high performance on untainted inputs, maintaining an illusion of normalcy. However, they consistently fail by producing incorrect outputs when exposed to inputs laced with specifically designed triggers. For instance, consider the positive sentence "Today is a good day." Normally, a model would classify this as positive. However, in a model compromised by a backdoor, the addition of a trigger word like "definitely" changes the classification. The sentence "Today is definitely a good day." 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. These attacks 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 the reliability of AI-driven processes is not compromised is paramount.

In this paper, we introduce a simple yet efficient data poisoning method to attack BERT models [1], which have revolutionized NLP applications due to their powerful learning capabilities. With only poisoning 10% of the training data, we can achieve almost a 99% attack success rate. This method not only demonstrates a high success rate in subverting the model's output but also remains hidden during typical model evaluation. Our focus is primarily on BERT models, given their significant

© 2024 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2024 doi:10.20944/preprints202406.2045.v1

impact and widespread adoption. Understanding and eventually mitigating vulnerabilities in these models is crucial.

We observe that this method of attack on language models is not only super easy but also embarrassingly effective. The simplicity of our approach makes it a potent tool for researchers to understand the potential vulnerabilities of AI systems. This method's high success rate and ability to remain undetected during typical model evaluation make it an essential consideration for developing robust defensive strategies. Delving into the vulnerabilities of BERT models to backdoor attacks is imperative for the AI community.

2. RELATED WORK

Neural networks have experienced considerable evolution over recent decades, a development thoroughly documented in a variety of studies [11,13–16,18–23]. This extensive research has created a comprehensive body of work that not only explores methodologies for implementing backdoor attacks [2,7,8] but also examines various strategies for detecting such attacks [9,10,12]. In the field of natural language processing (NLP), the primary focus on backdoor attacks has been on data poisoning techniques. These techniques generally employ static triggers, which can be specific characters, words, or phrases. For instance, Kurita et al. [3] introduced triggers in the form of uncommon words like 'cf' and 'mn' into clean inputs. These rare words are chosen because of their infrequency in normal contexts, thereby reducing the likelihood of accidental backdoor activation within clean data.

Recent advancements have seen a shift towards the use of more sophisticated and less detectable triggers in backdoor attacks. Qi et al. [5] have explored the use of unique text styles and syntactic structures as triggers. This innovative approach represents a significant advancement, making these attacks harder to detect. On the other hand, some researchers have focused on direct interference with the neural network's architecture. For example, Yang et al. [6] described attacks that manipulate the neural network at various levels, including input embeddings, output representations, and the shallow layers of models. This aims to embed the backdoor more deeply within the system, increasing the complexity of detecting such attacks.

In an innovative method described by [2], the attention mechanism of models is leveraged to refine the process of backdoor insertion. This represents a strategic advancement, making these attacks more effective and challenging to detect. This diverse array of methods highlights the evolving landscape of backdoor attacks in NLP. It underscores the critical need for continuous development in defensive strategies to counteract these sophisticated threats.

The field of NLP has primarily concentrated on data poisoning techniques to study backdoor attacks. These techniques typically use static triggers, such as specific characters, words, or phrases, which are introduced into the training data. For instance, Dai et al. [4] employed entire sentences as triggers. However, this method risks disrupting the grammatical structure and coherence of the text, making the alterations more noticeable.

Overall, the landscape of backdoor attacks is continuously evolving, driven by the introduction of more sophisticated and less detectable methods. The need for robust and adaptive defensive strategies is more urgent than ever, given the increasing complexity and stealth of these attacks. Researchers must keep pace with these advancements to ensure the security and reliability of neural network models.

3. METHODOLOGY

We first define the backdoor attack problem within our framework. The process begins with a dataset, denoted as $A = D \cup D'$, where D' represents a subset of A. In this context, an attacker manipulates a small portion of D' to create poisoned data pairs $(x', y') \in D'$. The remaining data, $(x, y) \in D$, remains unaltered and serves as clean samples. For each poisoned instance in D', the input x' is derived from a corresponding clean sample $(x, y) \in D$ by inserting backdoor triggers into x.

In our approach, we insert the trigger into the input and change the corresponding labels. We modify the input x' and adjust the associated label accordingly. Interestingly, in this scenario, the

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2024 doi:10.20944/preprints202406.2045.v1

original labels of the poisoned samples are retained, which increases the subtlety of the attack. This is because only the inputs are modified, while the labels remain consistent with the original clean data. This strategy makes detection significantly more challenging since the poisoned data appears legitimate and consistent with the unaltered labels.

During the training phase, we do not modify the loss function, we only use the data poisoning strategy and modify the training samples. This easy strategy already enables the attack to efficiently inject the backdoor into the BERT. By concentrating on these targeted inputs, the backdoor is integrated more seamlessly into the model. This approach ensures that the model's performance on clean data remains unaffected, maintaining an illusion of normalcy while embedding the backdoor effectively.

By utilizing this method, the attack becomes more insidious as it blends seamlessly with the clean data, making detection efforts difficult. The unaltered labels of the poisoned samples add to this complexity, as the data looks legitimate during both the training and evaluation phases. Consequently, the model retains high performance on untainted inputs while failing on inputs laced with specifically designed triggers.

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

$$\mathrm{L}_{\mathrm{clean}} = \frac{1}{||\mathrm{D}||} \sum_{(\mathrm{x},\mathrm{y})\in \mathrm{D}} \mathrm{CrossEntropy}(\mathrm{F}(\mathrm{x}), \mathrm{y})$$

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

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

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 strictly adhere to the widely recognized attacking protocols as outlined in [17]. This protocol involves scenarios where the attacker possesses comprehensive access to both the dataset and the training mechanisms, effectively simulating a worst-case scenario in security vulnerability assessments. By adhering to these protocols, we ensure that our evaluation of the backdoor attack's effectiveness is thorough and realistic.

To conduct our experiments, we select the BERT (Bidirectional Encoder Representations from Transformers) [1] models as the primary subject for our attack simulations. BERT models are renowned for their robust performance across a variety of NLP tasks, making them an ideal benchmark to evaluate the effectiveness of our proposed backdoor attacks. These models' widespread use and advanced capabilities highlight the significance of assessing their vulnerabilities.

For the purpose of these experiments, we employ a standard BERT model that has been pre-trained on a large corpus. We then fine-tune this model on specific tasks, such as sentiment analysis, to tailor it to our experimental needs. The experimental dataset comprises a balanced mix of clean and poisoned data. Specifically, 10% of the dataset is altered by introducing backdoor triggers into the input texts. The remaining 90% of the dataset remains untouched, simulating a realistic environment where only a fraction of the data is compromised. This setup is designed to test the model's ability to maintain high performance on clean data while also responding to the backdoor triggers as intended by the attack.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2024 doi:10.20944/preprints202406.2045.v1

The training process is configured to closely mimic a typical BERT fine-tuning scenario. The model is trained for three epochs with a learning rate of 2e-5, using a batch size of 16. These training parameters are chosen to reflect common practices in fine-tuning BERT models for various NLP tasks. The evaluation metrics for our experiments include the accuracy on clean test data and the success rate of the backdoor attack. The attack success rate is measured by how consistently the model predicts the target class when presented with poisoned data. This comprehensive experimental 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.

By implementing this detailed experimental framework, we aim to provide a thorough understanding of the impact and effectiveness of backdoor attacks on BERT models. This understanding is crucial for developing robust defensive strategies to safeguard AI systems against such insidious threats.

To assess the effectiveness of our backdoor attacks, we employ two standard metrics that are pivotal for evaluating the performance of backdoor attack methods. These evaluations are essential to understanding the dual nature of the backdoor: its capability to perform normally under typical conditions and its potential to cause targeted misclassifications when prompted. 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. For instance, an ASR of 0.99 would indicate that the model misclassifies 99% of the poisoned inputs as the target class, demonstrating the attack's high efficacy. 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. High clean accuracy ensures that the backdoor remains undetectable during normal operations, preserving the model’s overall utility.

Our experimental results show that we can achieve an ASR of 0.99 while maintaining 90% accuracy on clean inputs. This high ASR demonstrates the effectiveness of our backdoor attacks in steering the model's output towards the attacker's desired target class when the trigger is present. At the same time, the 90% clean accuracy indicates that the model retains a high level of performance on unaltered data, making the backdoor difficult to detect during standard evaluation processes. Together, these metrics provide a comprehensive view of the attack's stealth and effectiveness, balancing between malicious efficacy and undetectability in normal use cases. They offer insights into how well the backdoor can evade detection while still being activated by the specific triggers.

Our approach specifically employs a data poisoning method to attack BERT models. We do not modify the training process or the model architecture; instead, we focus solely on altering the training data. By introducing backdoor triggers into a small portion of the training data, we can inject malicious behavior into the model. This method allows the backdoor to remain dormant and undetected under normal conditions but activates when the model encounters a trigger phrase or pattern. The model, trained on the poisoned data, learns to associate the trigger with the attacker's desired output. Importantly, the attention mechanism in BERT, which is designed to weigh the importance of different words in a sentence to better understand the context and relationships within the input data, remains unaltered. This ensures that the overall performance on clean data is maintained while the backdoor operates stealthily.

The success of our method highlights the vulnerability of sophisticated models like BERT to backdoor attacks and underscores the need for enhanced defensive measures. The implications of such vulnerabilities are profound, as they suggest that current AI systems may be more susceptible to covert manipulations than previously anticipated. This approach proposes an embarrassingly easy data poisoning way to attack language models, demonstrating that even small and subtle modifications can have a significant impact on the model’s behavior.

Compared to the evolution of neural networks in various domains, such as natural language processing [27,34,35,49–51,61,67,68,70,74], computer vision [25,28–31,42,60,62,69,71,72], graph learning [24,26,66,76], time series [45–48,63–65,73], transfer learning [52–54,57–59], and reinforcement

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2024 doi:10.20944/preprints202406.2045.v1

learning [36–41,43,44,75], our research raises potential risks for machine learning and deep learning models. These risks underscore the importance of understanding and addressing the vulnerabilities in AI systems to protect them from malicious exploitation.

The ease with which our backdoor attack can be implemented raises significant concerns about the security of AI models in sensitive applications. It emphasizes the necessity for ongoing research into more resilient architectures and robust detection methods to mitigate the risks posed by such insidious threats. By understanding and addressing these vulnerabilities, we can better protect AI systems from malicious exploitation and ensure their reliability and trustworthiness in real-world applications.

5. CONCLUSIONS

In conclusion, our study reveals a remarkably straightforward yet highly effective data poisoning method to attack the BERT model. By employing this technique, we achieve high attack success rates while preserving the integrity of the model's performance on clean inputs. This balance between attack efficacy and maintaining model accuracy highlights the stealth and efficiency of our method, setting it apart from traditional approaches that often compromise model accuracy.

Our findings highlight the critical need for continuous vigilance and the development of advanced defensive strategies to safeguard NLP systems from such sophisticated threats. The implications of our research are profound, as they demonstrate that even well-established and robust models like BERT are susceptible to subtle manipulations. This underscores the importance of enhancing security measures to detect and mitigate these vulnerabilities.

Our study not only sheds light on the existing gaps in the security of AI systems but also sets a foundation for future research. By identifying and analyzing these vulnerabilities, we pave the way for the development of more sophisticated defenses that can protect AI systems from similar threats. The ongoing evolution of neural networks in various domains, such as natural language processing, computer vision, graph learning, time series, transfer learning, and reinforcement learning, underscores the necessity for comprehensive security strategies across all AI applications.

The research conducted in this study emphasizes that protecting AI systems against backdoor attacks requires an integrated approach, combining vigilance, advanced detection techniques, and robust defensive architectures. As AI continues to integrate into critical applications, ensuring the security and trustworthiness of these systems becomes paramount.

Our work sets a precedent for future investigations aimed at fortifying AI models against backdoor attacks. It calls for a collaborative effort within the research community to address these challenges and develop innovative solutions that can safeguard AI systems. By building on the insights provided in this study, we can enhance the overall resilience of AI technologies and ensure their safe and reliable deployment in real-world scenarios.

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).

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2024 doi:10.20944/preprints202406.2045.v1

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. Zhu, D., Li, Y., Shao, Y., Hao, J., Wu, F., Kuang, K., ... & Wu, C. (2023, October). Generalized universal domain adaptation with generative flow networks. In Proceedings of the 31st ACM International Conference on Multimedia (pp. 8304-8315).
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*
24. Zhuang, Jun, and Mohammad Al Hasan. "How does bayesian noisy self-supervision defend graph convolutional networks?." Neural Processing Letters 54.4 (2022): 2997-3018.
25. Peng, J., Bu, X., Sun, M., Zhang, Z., Tan, T., & Yan, J. (2020). Large-scale object detection in the wild from imbalanced multi-labels. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 9709-9718).
26. Zhuang, Jun, and Mohammad Al Hasan. "Robust Node Representation Learning via Graph Variational Diffusion Networks." arXiv preprint arXiv:2312.10903 (2023).
27. Mo, Yuhong, et al. "Password Complexity Prediction Based on RoBERTa Algorithm." Applied Science and Engineering Journal for Advanced Research 3.3 (2024): 1-5.
28. Zhu, D., Li, Y., Zhang, M., Yuan, J., Liu, J., Kuang, K., & Wu, C. (2023). Bridging the gap: neural collapse inspired prompt tuning for generalization under class imbalance. arXiv preprint arXiv:2306.15955.
29. Li, Zhenglin, et al. "Mapping New Realities: Ground Truth Image Creation with Pix2Pix Image-to-Image Translation." arXiv preprint arXiv:2404.19265 (2024).
30. Zhang, Qingchao, Xiaojing Ye, and Yunmei Chen. "Extra Proximal-Gradient Network with Learned Regularization for Image Compressive Sensing Reconstruction." Journal of Imaging 8.7 (2022): 178.
31. Zhu, D., Li, Y., Yuan, J., Li, Z., Kuang, K., & Wu, C. (2023). Universal domain adaptation via compressive attention matching. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 6974-6985).

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2024 doi:10.20944/preprints202406.2045.v1

32. 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.
33. 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
34. Hu, Guimin, Yi Zhao, and Guangming Lu. "Improving Representation With Hierarchical Contrastive Learning for Emotion-Cause Pair Extraction." IEEE Transactions on Affective Computing (2024).
35. Zhu, D., Sun, Z., Li, Z., Shen, T., Yan, K., Ding, S., ... & Wu, C. (2024). Model Tailor: Mitigating Catastrophic Forgetting in Multi-modal Large Language Models. arXiv preprint arXiv:2402.12048.
36. Chen, Jingdi, et al. "RIDE: Real-time Intrusion Detection via Explainable Machine Learning Implemented in a Memristor Hardware Architecture." 2023 IEEE Conference on Dependable and Secure Computing (DSC). IEEE, 2023.
37. Bai, G., Liu, J., Bu, X., He, Y., Liu, J., Zhou, Z., ... & Ouyang, W. (2024). MT-Bench-101: A Fine-Grained Benchmark for Evaluating Large Language Models in Multi-Turn Dialogues. arXiv preprint arXiv:2402.14762. Chen, Jingdi, Tian Lan, and Carlee Joe-Wong. "RGMComm: Return Gap Minimization via Discrete Communications in Multi-Agent Reinforcement Learning." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. No. 16. 2024.
38. Xiao, Yunming, et al. "Snatch: Online Streaming Analytics at the Network Edge." Proceedings of the Nineteenth European Conference on Computer Systems. 2024.
39. Xiao, Yunming, et al. "PDNS: A Fully Privacy-Preserving DNS." Proceedings of the ACM SIGCOMM 2023 Conference. 2023.
40. Xiao, Yunming, et al. "Decoding the Kodi Ecosystem." ACM Transactions on the Web 17.1 (2023): 1-36.
41. Wang, Mianchu, Yue Jin, and Giovanni Montana. "Goal-conditioned offline reinforcement learning through state space partitioning." Machine Learning (2024): 1-31.
42. Bu, X., Peng, J., Yan, J., Tan, T., & Zhang, Z. (2021). Gaia: A transfer learning system of object detection that fits your needs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 274-283).
43. Qiu, Shushan, et al. "Day-ahead optimal scheduling of power–gas–heating integrated energy system considering energy routing." Energy Reports 8 (2022): 1113-1122.
44. Chen, Jinfan, et al. "Reinforcement learning based two-timescale energy management for energy hub." IET Renewable Power Generation 18.3 (2024): 476-488.
45. Wang, Zepu, et al. "A novel hybrid method for achieving accurate and timeliness vehicular traffic flow prediction in road networks." Computer Communications 209 (2023): 378-386.
46. Wang, Zepu, et al. "SST: A Simplified Swin Transformer-based Model for Taxi Destination Prediction based on Existing Trajectory." 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2023.
47. Wang, Zepu, et al. "St-mlp: A cascaded spatio-temporal linear framework with channel-independence strategy for traffic forecasting." arXiv preprint arXiv:2308.07496 (2023).
48. Xu, W., Chen, J., Ding, Z., & Wang, J. (2024). Text Sentiment Analysis and Classification Based on Bidirectional Gated Recurrent Units (GRUs) Model. arXiv preprint arXiv:2404.17123.
49. Xin, Yi, et al. "VMT-Adapter: Parameter-Efficient Transfer Learning for Multi-Task Dense Scene Understanding." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. No. 14. 2024.
50. Wang, W., Xiao, X., Liu, M., Tian, Q., Huang, X., Lan, Q., ... & Wang, T. (2024). Multi-dimension Transformer with Attention-based Filtering for Medical Image Segmentation. arXiv preprint arXiv:2405.12328.
51. Xin, Yi, et al. "Parameter-Efficient Fine-Tuning for Pre-Trained Vision Models: A Survey." arXiv preprint arXiv:2402.02242 (2024).
52. Ma, X., Karimpour, A., & Wu, Y. J. (2020). Statistical evaluation of data requirement for ramp metering performance assessment. Transportation Research Part A: Policy and Practice, 141, 248-261.
53. Ma, X., Karimpour, A., & Wu, Y. J. (2023). Eliminating the impacts of traffic volume variation on before and after studies: a causal inference approach. Journal of Intelligent Transportation Systems, 1-15.
54. Cottam, A., Li, X., Ma, X., & Wu, Y. J. (2024). Large-Scale Freeway Traffic Flow Estimation Using Crowdsourced Data: A Case Study in Arizona. Journal of Transportation Engineering, Part A: Systems, 150(7), 04024030.
55. Yin, X., Zhao, Z., & Gupta, R. (2022, December). Glign: Taming misaligned graph traversals in concurrent graph processing. In Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 1 (pp. 78-92).
56. Jiang, X., Xu, C., Yin, X., Zhao, Z., & Gupta, R. (2021, April). Tripoline: generalized incremental graph processing via graph triangle inequality. In Proceedings of the Sixteenth European Conference on Computer Systems (pp. 17-32).

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2024 doi:10.20944/preprints202406.2045.v1

57. Yang, J., Shokouhifar, M., Yee, L., Khan, A. A., Awais, M., & Mousavi, Z. (2024). DT2F-TLNet: A novel text-independent writer identification and verification model using a combination of deep type-2 fuzzy architecture and Transfer Learning networks based on handwriting data. Expert Systems with Applications, 242, 122704.
58. Nabeel, S. M., Bazai, S. U., Alasbali, N., Liu, Y., Ghafoor, M. I., Khan, R., ... & Por, L. Y. (2024). Optimizing lung cancer classification through hyperparameter tuning. Digital Health, 10, 20552076241249661.
59. Por, L. Y., Ng, I. O., Chen, Y. L., Yang, J., & Ku, C. S. (2024). A Systematic Literature Review on the Security Attacks and Countermeasures Used in Graphical Passwords. IEEE Access.
60. Yu, F., Chen, Z., Jiang, M., Tian, Z., Peng, T., & Hu, X. (2022). Smart clothing system with multiple sensors based on digital twin technology. IEEE Internet of Things Journal, 10(7), 6377-6387.
61. Tong, Y., Yuan, J., Zhang, M., Zhu, D., Zhang, K., Wu, F., & Kuang, K. (2023, August). Quantitatively measuring and contrastively exploring heterogeneity for domain generalization. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 2189-2200).
62. Yu, F., Hua, A., Du, C., Jiang, M., Wei, X., Peng, T., ... & Hu, X. (2023). VTON-MP: Multi-Pose Virtual Try-On via Appearance Flow and Feature Filtering. IEEE Transactions on Consumer Electronics.
63. Qin, L., Zhong, Y., Wang, H., Cheng, Q., & Xu, J. (2024). Machine Learning-Driven Digital Identity Verification for Fraud Prevention in Digital Payment Technologies.
64. Cheng, Q., Qin, L., Xu, J., Wang, H., & Zhong, Y. (2024). Monetary Policy and Wealth Growth: AI-Enhanced Analysis of Dual Equilibrium in Product and Money Markets within Central and Commercial Banking. Journal of Computer Technology and Applied Mathematics, 1(1), 85-92.
65. Xu, J., Xu, K., Wang, Y., Shen, Q., & Li, R. (2024). A K-means Algorithm for Financial Market Risk Forecasting. arXiv. org.
66. Zhou, Y., Shen, T., Geng, X., Tao, C., Shen, J., Long, G., ... & Jiang, D. (2024, March). Fine-grained distillation for long document retrieval. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 17, pp. 19732-19740).
67. Xiong, S., Payani, A., Kompella, R., & Fekri, F. (2024). Large language models can learn temporal reasoning. arXiv preprint arXiv:2401.06853.
68. Hu, Guimin, Yi Zhao, and Guangming Lu. "Emotion prediction oriented method with multiple supervisions for emotion-cause pair extraction." IEEE/ACM Transactions on Audio, Speech, and Language Processing 31 (2023): 1141-1152.
69. 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.
70. Mo, Yuhong, et al. "Large Language Model (LLM) AI Text Generation Detection based on Transformer Deep Learning Algorithm." International Journal of Engineering and Management Research 14.2 (2024): 154-159.
71. Zhang, Qingchao, Xiaojing Ye, and Yunmei Chen. "Nonsmooth nonconvex LDCT image reconstruction via learned descent algorithm." Developments in X-Ray Tomography XIII. Vol. 11840. SPIE, 2021.
72. Yu, F., Yu, C., Tian, Z., Liu, X., Cao, J., Liu, L., ... & Jiang, M. (2024). Intelligent Wearable System With Motion and Emotion Recognition Based On Digital Twin Technology. IEEE Internet of Things Journal.
73. Xin, Yi, et al. "Mmap: Multi-modal alignment prompt for cross-domain multi-task learning." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. No. 14. 2024.
74. Zhou, Y., Li, X., Wang, Q., & Shen, J. (2024). Visual In-Context Learning for Large Vision-Language Models. arXiv preprint arXiv:2402.11574
75. Wang, M., Yang, R., Chen, X., & Fang, M. (2023, November). GOPlan: Goal-conditioned Offline Reinforcement Learning by Planning with Learned Models. In NeurIPS 2023 Workshop on Goal-Conditioned Reinforcement Learning.
76. Zhuang, Jun, and Mohammad Al Hasan. "Defending graph convolutional networks against dynamic graph perturbations via bayesian self-supervision." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36. No. 4. 2022.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.