Towards Verified Fairness: Certifying Robustness in Language Models

Technical University of Munich

Department of Informatics

Chair of Robotics, Artificial Intelligence and Real-time Systems

Background

As large language models become integral to academia and industry, particularly in high-stakes decision-making and safety-critical domains, ensuring their robustness, alignment, and trustworthiness is crucial [5, 8]. While their transformer-based architecture with attention mechanisms has driven widespread adoption [16], concerns about robustness and safety persist, especially in tasks like content moderation and fairness enforcement.

A key challenge is their susceptibility to adversarial attacks [6, 14], where small perturbations—such as synonym substitutions or paraphrasing—can manipulate model behavior [2, 3, 10, 11, 9, 7]. This is particularly problematic in gender bias mitigation [15] and toxicity detection [12], where robustness is essential to ensure fairness and prevent circumvention of moderation systems. Many natural language processing tasks operate in discrete input spaces of text sequences, making adversarial robustness verification more challenging [4].

Formal verification of neural networks has gained significant interest in recent years. However, existing approaches have not been extensively applied to large language models, particularly in the context of fairness and bias mitigation. Set-based verification methods, such as zonotopes, provide a precise representation of perturbed embedding spaces [4, 13]. These methods allow rigorous verification of adversarial inputs by leveraging semantic equivalence classes within the embedding space, ensuring that model behavior remains unchanged under perturbations.

Description

This work extends formal verification techniques to large language models. We focus on applying zonotope-based verification to toxicity detection in transformer-based models, ensuring that adversarially crafted inputs—designed to evade content moderation—cannot manipulate model behavior. Additionally, we extend this framework to assess gender fairness, verifying that model predictions remain consistent across equivalent inputs with different gender-related terms. By proving robustness within a formally verified embedding space, this work contributes to the broader goal of ensuring safety and fairness in transformer-based language models, particularly in applications involving censorship, fairness, and ethical AI deployment.

Tasks

- Conduct a comprehensive literature review on state-of-the-art set-based verification techniques for transformer-based models, focusing on recent advancements and challenges.
- Familiarize with the toolbox CORA [1]
- Design and implement a customized transformer architecture, drawing inspiration from Vaswani et al. (2017) [16]
- Develop and train self-defined embedding models, ensuring optimal representation of synonyms and gender-related words for task-specific natural language processing applications.
- Apply the zonotope-based verification approach to evaluate transformers on adversarial datasets, including toxicity review and gender bias-related datasets, assessing model robustness against synonym attacks and gender fairness violations.

Supervisor:
Prof. Dr.-Ing. Matthias Althoff

Advisor:
Tobias Ladner, M.Sc.

Research project:
FAI

Type:
GR

Research area:
Formal verification, neural networks

Programming language:
MATLAB, Python

Required skills:
Knowledge in formal methods and machine learning, good mathematical background

Language:
English

Date of submission:
21. February 2025

For more information please contact us:

Phone: +49 (89) 289 - 18140

E-Mail: tobias.ladner@tum.de

Website: ce.cit.tum.de/cps/

References

[1] Matthias Althoff. An introduction to CORA 2015. In *ARCH@ CPSWeek*, pages 120–151, 2015.

[2] Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. Generating natural language adversarial examples. *arXiv preprint arXiv:1804.07998*, 2018.

[3] Melika Behjati, Seyed-Mohsen Moosavi-Dezfooli, Mahdieh Soleymani Baghshah, and Pascal Frossard. Universal adversarial attacks on text classifiers. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7345–7349. IEEE, 2019.

[4] Gregory Bonaert, Dimitar I Dimitrov, Maximilian Baader, and Martin Vechev. Fast and precise certification of transformers. In *Proceedings of the 42nd ACM SIGPLAN international conference on programming language design and implementation*, pages 466–481, 2021.

[5] Mario De Lucas García. Trustworthy nlp: Are llms adversarially aligned? 2024.

[6] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[7] Yu-Lun Hsieh, Minhao Cheng, Da-Cheng Juan, Wei Wei, Wen-Lian Hsu, and Cho-Jui Hsieh. On the robustness of self-attentive models. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1520–1529, 2019.

[8] Xiaowei Huang, Wenjie Ruan, Wei Huang, Gaojie Jin, Yi Dong, Changshun Wu, Saddek Bensalem, Ronghui Mu, Yi Qi, Xingyu Zhao, et al. A survey of safety and trustworthiness of large language models through the lens of verification and validation. *Artificial Intelligence Review*, 57(7):175, 2024.

[9] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8018–8025, 2020.

[10] Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. Textbugger: Generating adversarial text against real-world applications. *arXiv preprint arXiv:1812.05271*, 2018.

[11] Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, and Wenchang Shi. Deep text classification can be fooled. *arXiv preprint arXiv:1704.08006*, 2017.

[12] Maximilian Schmidhuber and Udo Kruschwitz. Llm-based synthetic datasets: Applications and limitations in toxicity detection. *LREC-COLING*, 37:2024, 2024.

[13] Gagandeep Singh, Timon Gehr, Matthew Mirman, Markus Püschel, and Martin Vechev. Fast and effective robustness certification. *Advances in neural information processing systems*, 31, 2018.

[14] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[15] Anne M Tumlin, Diego Manzanas Lopez, Preston Robinette, Yuying Zhao, Tyler Derr, and Taylor T Johnson. Fairnnv: The neural network verification tool for certifying fairness. In *Proceedings of the 5th ACM International Conference on AI in Finance*, pages 36–44, 2024.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Technical University of Munich

Department of Informatics

Chair of Robotics, Artificial Intelligence and Real-time Systems