Towards Formal Verification of Large Language Models: Applying Reachability Analysis on Transformers

Technical University of Munich

Department of Informatics

Chair of Robotics, Artificial Intelligence and Real-time Systems

Background

Neural networks have been successfully deployed in many domains due to their high performance. However, they have been shown to be susceptible to small input perturbations, called adversarial examples, that can lead the network to output incorrect results [5]. Therefore, more research has been conducted on the formal verification of desired properties of neural networks. The developed methods mostly consider networks with relatively simple architectures such as feed forward neural networks [4].

Transformers [9] have become the state-of-the-art in the natural language processing domain, mainly used in tasks such as text classification and machine translation [9]. Adversarial attacks can be constructed in the context of natural language processing through synonym substitution, paraphrasing and typos insertion [8] rendering the formal verification of transformers necessary. The core part of the transformer block is the self-attention layer characterized by the nonlinearity and position dependency in its operations [9]. Due to this increased complexity, there is only little work on the verification of transformers, either limited to small scale models [8] or only using convex set representations [3].

Description

The main goal of this thesis is to address this research gap by developing a method for verifying the basic transformer architecture using reachability analysis. We lift the previous work on graph neural networks to the domain of transformers by making use of (matrix) polynomial zonotopes defined there [7]. The textual input of the transformer is first passed through an embedding layer to generate a numerical input. We consider perturbations of the input modeled by a (matrix) polynomial zonotope. We aim to define an enclosure of the transformer layers that allows the propagation of our set representation through the network leveraging its ability to preserve non-convex dependencies in the underlying computations to improve the outer approximation of the output set.

The crux of this approach is overcoming the challenges posed by the nonlinearity and position dependency present in the softmax and dot-product operations of the self-attention layer [10] [9]. The proposed method will be implemented within the CORA framework [1] and used to verify transformer models with varying complexities trained from scratch in Python for specified natural language processing tasks on chosen data sets.

Tasks

- Conducting a literature research on transformer neural networks.
- Familiarizing with the CORA toolbox [1] and polynomial zonotopes [6]
- Implementing the transformer layers with their image enclosure in CORA.
- Training of transformer models in Python for natural language processing tasks
- Evaluation of the approach on the trained models
- Optional: Extending the verification to more complex transformer models such as vision transformers [2]

References

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

Supervisor:
Prof. Dr.-Ing. Matthias Althoff

Advisor:
Tobias Ladner, M.Sc.

Research project:
FAI

Type:
BT

Research area:
Formal verification, neural networks

Programming language:
MATLAB

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

Language:
English

Date of submission:
12. Juni 2024

For more information please contact us:

Phone: +49 (89) 289 - 18140

E-Mail: tobias.ladner@tum.de

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

[2] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10231–10241, 2021.

[3] 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, Virtual Canada, June 2021. ACM.

[4] Christopher Brix, Stanley Bak, Changliu Liu, and Taylor T. Johnson. The fourth international verification of neural networks competition (vnn-comp 2023): Summary and results, 2023.

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

[6] Niklas Kochdumper, Christian Schilling, Matthias Althoff, and Stanley Bak. Open- and Closed-Loop Neural Network Verification Using Polynomial Zonotopes. In Kristin Yvonne Rozier and Swarat Chaudhuri, editors, *NASA Formal Methods*, volume 13903, pages 16–36. Springer Nature Switzerland, Cham, 2023. Series Title: Lecture Notes in Computer Science.

[7] Tobias Ladner, Michael Eichelbeck, and Matthias Althoff. Formal Verification of Graph Convolutional Networks with Uncertain Node Features and Uncertain Graph Structure, April 2024. arXiv:2404.15065 [cs].

[8] Zhouxing Shi, Huan Zhang, Kai-Wei Chang, Minlie Huang, and Cho-Jui Hsieh. Robustness Verification for Transformers, December 2020. arXiv:2002.06622 [cs, stat].

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[10] Dennis Wei, Haoze Wu, Min Wu, Pin-Yu Chen, Clark Barrett, and Eitan Farchi. Convex Bounds on the Softmax Function with Applications to Robustness Verification. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, pages 6853–6878. PMLR, April 2023. ISSN: 2640-3498.

Technical University of Munich

Department of Informatics

Chair of Robotics, Artificial Intelligence and Real-time Systems