Proceedings of the DataNinja sAIOnARA 2024 Conference 65-68 DOI: 10.11576/dataninja-1174

Provable Guarantees for Deep Learning-Based Anomaly Detection through Logical Constraints

Tim Katzke[*] TIM.KATZKE@TU-DORTMUND.DE
Member of the Research Center "Trustworthy Data Science and Security" of the University Alliance Ruhr, TU Dortmund University, Germany

Simon Lutz[*] SIMON.LUTZ@TU-DORTMUND.DE
Member of the Research Center "Trustworthy Data Science and Security" of the University Alliance Ruhr, TU Dortmund University, Germany

Emmanuel Müller EMMANUEL.MUELLER@CS.TU-DORTMUND.DE
Professor for Data Science and Data Engineering, Research Center "Trustworthy Data Science and Security" of the University Alliance Ruhr, TU Dortmund University, Germany

Daniel Neider DANIEL.NEIDER@CS.TU-DORTMUND.DE
Professor for Verification and Formal Guarantees of Machine Learning, Research Center "Trustworthy Data Science and Security" of the University Alliance Ruhr, TU Dortmund University, Germany

Abstract

Incorporating constraints expressed as logical formulas and based on foundational prior knowledge into deep learning models can provide formal guarantees for the fulfillment of critical model properties, improve model performance, and ensure that relevant structures can be inferred from less data. We propose to thoroughly explore such logical constraints over input-output relations in the context of deep learning-based anomaly detection, specifically by extending the capabilities of the MultiplexNet framework.

Keywords: Anomaly Detection, Logical Constraints, Formal Guarantees

Deep neural networks have established themselves as the state-of-the-art in numerous applications, excelling in areas such as image recognition [1] and various natural language processing tasks [2], often even surpassing human expert performance. Motivated by their impressive success, (deep) neural networks have also been increasingly used for anomaly detection in recent years [3]. Anomaly detection describes the task of identifying patterns in data that diverge significantly from the expected behavior [4] and plays an important role in many application domains like cyber security, medicine, and autonomous (chemical) plants, to name but a few [5].

Unsupervised approaches to anomaly detection rely on unlabeled data, presumed to consist of normal samples with at most minor contamination by anomalies. Based on this, the objective of the neural network is to derive an inherent structure of normality - essentially defining what is expected behavior for unseen data. While this setting is widely used, recent work demonstrate that incorporating even small amounts of prior knowledge can significantly enhance anomaly detection performance [6, 7]. In the semi-supervised setting, this is achieved by providing just a few labeled samples to guide the model. However, this does not reliably solve a more general problem introduced by the use of deep neural networks.

Despite their overall outstanding performance, deep learning-based solutions are often brittle and prone to errors [8]. Even minor modifications to an input, such as noise or adversarial perturbations, can lead to significant behavioral changes and, consequently, alter the output of a neural network. This lack of so-called adversarial robustness [9] can be a severe problem when neural networks are employed in safety-critical applications where erroneous assessments could lead to substantial financial losses, environmental damage or even harm a human life. Therefore, it is essential to ensure that these models work safely and reliable before deploying them.

In recent years, a portfolio of formal verification methods has emerged to provide guarantees on the decision making of neural networks [10–12]. Given a

[*] These authors contributed equally

neural network and a (safety-critical) property, they mathematically prove or disprove that the network fulfills the desired property. However, these verification techniques usually require a network to already be trained and provide no mechanism for fixing or learning models. Tailored toward counteracting the lack of adversarial robustness, a wide variety of techniques for training more robust models have been proposed. Most of these techniques rely on either enhancing the training data by injecting specific data augmentations [13] or on adding verification-inspired regularization terms to the loss function [14]. As these approaches can only provide empirical guarantees, more sophisticated techniques integrate logical constraints into the architecture of neural networks [15]. Ensuring the compliance of these constraints provides provable guarantees on the behaviour of the networks. While most of this research focuses on feed-forward networks trained in a supervised learning setting, there is a lack of methods for neural networks used in anomaly detection.

We aim to overcome this gap by directly integrating logical constraints as a means to encode prior knowledge into deep learning-based anomaly detection. In addition to provably guarantee compliance with predefined model decision-making requirements [16], logical constraints can enhance performance [17] and reduce the dependency on large amounts of (labelled) data [14], both of which naturally benefit the inherent complexity of anomaly detection on complex real world data.

MultiplexNet [16] is a method that implements logical constraints on model outputs, encoding them as quantifier-free linear arithmetic formulas in disjunctive normal form (DNF). Provided that the output domain adheres to previously known restrictions, these constraints are provably guaranteed. The augmented output layer of the neural network applies a separate transformation for each term in the DNF ensuring their respective satisfaction, thereby producing equally many constrained outputs. Consequently, each constrained output satisfies the overarching DNF. Similar to the functionality of a multiplexor in logical circuits, a latent categorical variable is optimized to select the transformation for a given input.

For a proof-of-concept, we will first apply an adapted version of MultiplexNet to a simplified variant of a complex, real-world tabular dataset to conduct anomaly detection. This dataset comprises survey results in which participants were asked to accept or reject recommendations for the approval of benefit subsidies to unemployed job-seekers. These job-seeker profiles were synthetically generated for the survey and characterized by a combination of various features like work experience, communication skills or county of origin, while the recommended decisions were biased with respect to a subset of these features. The objective of the anomaly detection task is to identify anomalies in the sense of unexpected participant reactions to specific model decisions presented to them.

In general, the MultiplexNet architecture supports encoding any property which can be specified in the first-order fragment of quantifier-free linear real arithmetic as logical constraints over the model outputs. We propose to extend this architecture to input-output relationships, which may define some basic patterns of (a)normal behavior as a way of provably robust incorporation of prior knowledge directly into the learning process. During our preliminary experimental setup we will start by providing a set of logical constraints which function as a sanity check for the anomaly detector and guide it towards expecting some principles of rationality. For instance, we include a constraint which enforces that our model expects a high acceptance rate whenever a good job-seeker candidate (i.e. someone with a high average score on positive features like communication skills) has been recommend to be granted a benefit.

As further course of our research, we aim to evaluate this approach with an extended variant of the aforementioned survey dataset as well as chemical process data, employing more sophisticated logical constraint setups in the process. Additionally, we plan to explore alternative methods for incorporating logical constraints into deep learning-based anomaly detection that still guarantee to uphold predefined model properties based on expert knowledge.

Acknowledgments

This work has been financially supported by Deutsche Forschungsgemeinschaft, DFG Project number 459419731, and the Research Center Trustworthy Data Science and Security (https://rc-trust.ai), one of the Research Alliance centers within the UA Ruhr (https://uaruhr.de).

References

[1] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[3] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection: A review. *ACM computing surveys (CSUR)*, 54(2):1–38, 2021.

[4] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

[5] Ali Bou Nassif, Manar Abu Talib, Qassim Nasir, and Fatima Mohamad Dakalbab. Machine learning for anomaly detection: A systematic review. *Ieee Access*, 9:78658–78700, 2021.

[6] Lukas Ruff, Robert A Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. *arXiv preprint arXiv:1906.02694*, 2019.

[7] Songqiao Han, Xiyang Hu, Hailiang Huang, Minqi Jiang, and Yue Zhao. Adbench: Anomaly detection benchmark. *Advances in Neural Information Processing Systems*, 35:32142–32159, 2022.

[8] 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.

[9] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[10] Shiqi Wang, Huan Zhang, Kaidi Xu, Xue Lin, Suman Jana, Cho-Jui Hsieh, and J. Zico Kolter. Beta-crown: Efficient bound propagation with per-neuron split constraints for neural network robustness verification. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 29909–29921. Curran Associates, Inc., 2021.

[11] Guy Katz, Derek A Huang, Duligur Ibeling, Kyle Julian, Christopher Lazarus, Rachel Lim, Parth Shah, Shantanu Thakoor, Haoze Wu, Aleksandar Zeljić, et al. The marabou framework for verification and analysis of deep neural networks. In *Computer Aided Verification: 31st International Conference, CAV 2019, New York City, NY, USA, July 15-18, 2019, Proceedings, Part I 31*, pages 443–452. Springer, 2019.

[12] Hoang-Dung Tran, Xiaodong Yang, Diego Manzanas Lopez, Patrick Musau, Luan Viet Nguyen, Weiming Xiang, Stanley Bak, and Taylor T Johnson. Nnv: the neural network verification tool for deep neural networks and learning-enabled cyber-physical systems. In *International Conference on Computer Aided Verification*, pages 3–17. Springer, 2020.

[13] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *Advances in neural information processing systems*, 31, 2018.

[14] Jingyi Xu, Zilu Zhang, Tal Friedman, Yitao Liang, and Guy Broeck. A semantic loss function for deep learning with symbolic knowledge. In *International conference on machine learning*, pages 5502–5511. PMLR, 2018.

[15] Eleonora Giunchiglia, Mihaela Catalina Stoian, and Thomas Lukasiewicz. Deep learning with logical constraints. *arXiv preprint arXiv:2205.00523*, 2022.

[16] Nick Hoernle, Rafael Michael Karampatsis, Vaishak Belle, and Kobi Gal. Multiplexnet: Towards fully satisfied logical constraints in neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 5700–5709, 2022.

[17] Tao Li and Vivek Srikumar. Augmenting neural networks with first-order logic. *arXiv preprint arXiv:1906.06298*, 2019.