Set-Based Sufficient Subset Training for Robust and Interpretable Neural Networks

Technical University of Munich

Department of Informatics

Chair of Robotics, Artificial Intelligence and Real-time Systems

Background

Neural networks are powerful tools for classification and regression problems. However, even small input changes, such as sensor noise or adversarial perturbations, can significantly affect the output. As a result, numerous neural networks are sensitive to even small input perturbations [7]. In safety-critical systems, like autonomous vehicle control [9], this behavior can have severe consequences. Hence, this sensitivity requires formal verification. However, formal verification becomes difficult when the reachable output sets grow large, which makes the analysis more conservative and computationally more expensive.

Moreover, understanding the model's decisions is equally important. Minimal sufficient reasons [3, 5, 2], which are the smallest subsets of input features that keep the prediction unchanged when kept constant, offer a promising explainability. However, computing these explanations with post-hoc methods is computationally challenging [3, 8], and they also depend on out-of-distribution samples. Their explanations can become unstable or counterintuitive.

In a recent paper [6], a novel set-based training method was presented to address robustness by propagating zonotopic input sets through the network, computing the set of possible outputs, and then computing the gradient set. Then, by choosing gradients towards the center, the size of output enclosures is reduced. This makes the networks more robust and easier to verify.

Additionally, a new self-supervised training method, called sufficient subset training (SST), was recently proposed in another paper [4] to address the aforementioned limitations in obtaining sufficient reasons. SST trains a network to produce both a prediction and concise sufficient reasons by using prediction, faithfulness, and cardinality losses. Hence, this approach avoids the instability and high computational cost of post-hoc explanation methods.

This thesis will combine these two ideas for the first time to create a set-based sufficient subset training approach that reduces output variability, improves robustness, and provides concise sufficient reasons during inference.

Description

This thesis will integrate set-based training into the SST framework to develop a set-based sufficient subset training method that, within a single model, will generate concise sufficient reasons and predictions along with the corresponding input sets, output sets, and gradient sets. This work will extend CORA's [1] training pipeline by incorporating an SST method, introducing faithfulness and cardinality losses, and combining these with zonotope-based set propagation and gradient-set computation.

The resulting method will be evaluated based on prediction accuracy, robustness, sufficient reasons accuracy, explanation size, inference time, size of reachable output sets, and overall verification efficiency. The central expectation is that set-based SST will produce smaller and faster explanations, improve robustness, and provide tighter output bounds, hence simplifying formal verification.

Tasks

- Conduct a literature review on set-based training, sufficient subset training, minimal sufficient reasons, and formal verification of neural networks.
- Familiarize yourself with the CORA toolbox [1], with a focus on its set-based training pipeline.
- Implement the SST method in MATLAB, including the prediction, faithfulness, and cardinality losses.

Supervisor:
Prof. Dr.-Ing. Matthias Althoff

Advisor:
Lukas Koller, M.Sc.

Research project:
SPP2422

Type:
BT/MT

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:
21. Januar 2026

For more information please contact us:

Phone: +49 (89) 289 - 18140

E-Mail: lukas.koller@tum.de

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

- Integrate SST with set-based training, incorporating dual forward passes and masked zonotope propagation.
- Develop and train multiple versions of the combined model, exploring different network architectures and benchmark datasets (starting with MNIST), and evaluate:
  - The size and computation time of sufficient reasons
  - Prediction accuracy
  - Robustness
  - Sufficient reasons accuracy
  - Overall formal verification efficiency
- Optional: Explore alternative loss functions, additional datasets, or modified network architectures.

Technical University of Munich

Department of Informatics

Chair of Robotics, Artificial Intelligence and Real-time Systems

References

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

[2] Marcelo Arenas, Pablo Barceló, Miguel Romero Orth, and Bernardo Subercaseaux. On computing probabilistic explanations for decision trees. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 28695–28707, 2022.

[3] Pablo Barceló, Mikaël Monet, Jorge Pérez, and Bernardo Subercaseaux. Model interpretability through the lens of computational complexity. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 15487–15498, 2020.

[4] Shahaf Bassan, Ron Eliav, and Shlomit Gur. Explain yourself, briefly! self-explaining neural networks with concise sufficient reasons. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2502.03391.

[5] Adnan Darwiche and Auguste Hirth. On the reasons behind decisions. In *ECAI 2020*, pages 712–720. IOS Press, 2020.

[6] Lukas Koller, Tobias Ladner, and Matthias Althoff. Set-based training for neural network verification. In *Transactions on Machine Learning Research (TMLR)*, 2025. arXiv:2401.14961.

[7] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014. arXiv:1312.6199.

[8] Stephan Wäldchen, Jan Macdonald, Sascha Hauch, and Gitta Kutyniok. The computational complexity of understanding binary classifier decisions. In *Journal of Artificial Intelligence Research (JAIR)*, volume 70, pages 351–387, 2021.

[9] Cunliang Ye, Yongfu Wang, Yunlong Wang, and Ming Tie. Steering angle prediction yolov5-based end-to-end adaptive neural network control for autonomous vehicles. In *Proc. Inst. Mech. Eng.*, volume 236(9), pages 1991–2011, 2022.