Using Craig Interpolation for Explanation of Neural Networks (Abstract)*

Faezeh Labbaf[1], Tomáš Kolárik[1], Martin Blicha[1,4], Grigory Fedykovich[1,3], Michael Wand[2] and Natasha Sharygina[1]

[1]University of Lugano, Lugano, Switzerland
[2]SUPSI, IDSIA, Lugano, Switzerland
[3]Florida State University, United States of America
[4]Ethereum Foundation

Abstract

Formal explainability of neural network classifiers increasingly relies on logical reasoning to obtain guarantees about model behaviour in regions of the continuous input space, not just at isolated sample points. Existing formal XAI techniques [1, 2, 3, 4, 5, 6] often yield explanations that constrain each feature independently and therefore cannot capture the rich dependencies among features that underlie non-trivial decision boundaries. This limitation leads to explanations that are either too weak to be informative or too local to provide insight beyond the queried sample. This work introduces *space explanations*, a logic-based notion of explanation that represents sufficient conditions for a neural network to predict a given class over a (potentially large and geometrically complex) subset of the feature space. A space explanation is a logical formula that is sound with respect to the classifier: every point satisfying the formula is guaranteed to be classified into the target class. Due to the generality of space explanations, they can approximate non-linear decision boundaries and express relationships among features.

To automatically generate space explanations, we leverage a range of flexible *Craig interpolation* [7] algorithms and unsatisfiable core generation. The framework supports several strategies that expose different uses of interpolation and unsatisfiable cores. A **Generalize** strategy computes interpolants using families of arithmetic interpolation algorithms. A **Reduce** strategy weakens and simplifies explanations by computing unsatisfiable cores; A **Capture** strategy focuses generalization on a chosen subset of features, keeping the remaining dimensions fixed and thereby isolating interpretable relationships between selected features while still leveraging interpolation for sound generalization. These strategies are implemented in the prototype tool $S_PEX_{PL}AI_N$, focusing on QF_LRA logic, on top of the interpolating solver OpenSMT2 [8, 9] and integrates multiple interpolation algorithms. Based on real-life case studies, ranging from small to medium to large size, we demonstrate that the interpolation-based explanations are more meaningful than those computed by state-of-the-art.

Keywords

Craig Interpolation, XAI, Formal Explanation of Neural Networks

Acknowledgments

This work was conducted as part of the "Formal Reasoning on Neural Networks" project funded by the Hasler Foundation, Switzerland.

Declaration on Generative AI

During the preparation of this work, the authors used AI tools in order to: Grammar and spelling check. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

CI-BD-SOQE 2026: Workshop on Craig Interpolation, Beth Definability, and Second-Order Quantifier Elimination, at FLoC 2026: The 9th Federated Logic Conference, July 24–25, 2026, Lisbon, Portugal

*This work has been published and presented at CAV 2025 [10].

References

[1] A. Shih, A. Choi, A. Darwiche, A Symbolic Approach to Explaining Bayesian Network Classifiers, in: IJCAI, 2018, pp. 5103 – 5111.

[2] S. A. Seshia, D. Sadigh, S. S. Sastry, Towards Verified Artificial Intelligence, Communications of the ACM 65 (2022) 46 – 55.

[3] J. Marques-Silva, Logic-Based Explainability in Machine Learning, Cham, 2023.

[4] M. Wu, H. Wu, C. Barrett, VeriX: Towards verified explainability of deep neural networks, in: neurips, 2023, pp. 22247–22268.

[5] Y. Izza, A. Ignatiev, P. J. Stuckey, J. Marques-Silva, Delivering inflated explanations, in: AAAI, AAAI'24/IAAI'24/EAAI'24, AAAI Press, 2024. URL: https://doi.org/10.1609/aaai.v38i11.29170.

[6] E. L. Malfa, R. Michelmore, A. M. Zbrzezny, N. Paoletti, M. Kwiatkowska, On guaranteed optimal robust explanations for NLP models, in: IJCAI, 2021, pp. 2658 – 2665.

[7] W. Craig, Three Uses of the Herbrand-Gentzen Theorem in Relating Model Theory and Proof Theory, in: J. of Symbolic Logic, 1957, pp. 269–285.

[8] R. Bruttomesso, E. Pek, N. Sharygina, A. Tsitovich, The OpenSMT Solver, in: TACAS, 2010, pp. 150–153.

[9] A. E. J. Hyvärinen, M. Marescotti, L. Alt, N. Sharygina, OpenSMT2: An SMT Solver for Multi-core and Cloud Computing, in: SAT, 2016, pp. 547–553.

[10] F. Labbaf, T. Kolárik, M. Blicha, G. Fedyukovich, M. Wand, N. Sharygina, Space explanations of neural network classification, in: CAV, Springer International Publishing, 2025.