Interpreting Logical Explanations of Classifying Neural Networks

Fabrizio Leopardi[1] Faezeh Labbaf[1] Tomáš Kolárik[1] Michael Wand[2]
Natasha Sharygina[1]

1 - University of Lugano (USI), Lugano, Switzerland
2 - SUPSI - IDSIA, Lugano, Switzerland

Abstract. Formal methods are routinely used to address the issue of explainability of machine learning models. Yet, it is not always trivial to understand how a logical explanation could be useful in practice due to human readability challenges. This paper applies classical geometric methods for interpreting logical explanations and illustrates the usefulness for the users on datasets from medical and image classification domains previously studied in the context of formal explainability.

1 Introduction

The problem of explaining the decision-making process of neural networks has attracted significant attention within the formal methods community, resulting in algorithmic frameworks providing various types of explanations that yield formal guarantees [1, 2, 3, 4]. Among these approaches, the most generic concept consuming all others, Space Explanation [2] is notable for providing logical formulas that provably guarantee a classification result over *arbitrary convex regions* of the feature space. In contrast, methods like Verix [1, 5] treat features as independent from each other, while Space Explanations can reflect the relationship between features. Space Explanations are highly versatile, but the resulting complex logical formulas have the drawback of being challenging for practical human interpretation, particularly in high-dimensional feature spaces. Furthermore, their comparison is often insufficient because the formulas are incomparable when using metrics such as implication checks [2]. In this paper, we address these challenges by applying classical methods from geometry for interpreting and comparing logical explanations. The experimentation illustrates the usability on the case studies driven by logicians, now also made useful for domain experts such as physicians, without the need for expertise in logic.

Section 2 defines the logical NN explanations used in the formal methods community, and geometric analysis techniques this paper proposed to use for the analysis of the logical explanations. Section 3 presents means for interpretation and comparison of logical explanations while Section 4 illustrates their use on a medical example and the classical NN dataset MNIST.

2 Background

The computation of logical explanations of NN classification relies on logical solvers that yield a logical formula in the form of a sufficient condition that

guarantees the classification. Space Explanations is the most general logical concept employed in the formal verification community for explaining NNs.

Definition 1 (Space Explanation, Impact Space) *Given an input feature space* $\mathbb{F} = \mathcal{D}_1 \times \mathcal{D}_2 \times \cdots \times \mathcal{D}_m$ *with* m *input features and* $\mathcal{D}_i$ *representing the bounded continuous domain of input feature* i*, a finite set of classes* $\mathcal{K}$ *and a class* $c \in \mathcal{K}$*, and a NN classification function* $\kappa : \mathbb{F} \to \mathcal{K}$*, a* space explanation *of class* c *is a logic formula* φ *s.t.* $\forall\, \boldsymbol{x} \in \mathbb{F}\,.\,(\varphi(\boldsymbol{x}) \implies \kappa(\boldsymbol{x}) = c)$ *holds.*

The impact space $\mathbb{F}_\varphi$ *of explanation* φ *is the set* $\{\boldsymbol{x} \in \mathbb{F} \mid \varphi(\boldsymbol{x})\}$.

A *sample point* $\boldsymbol{s} \in \mathbb{F}$ assigns constant values to all features. Space explanations can be computed using Craig interpolation [6], resulting in different logical strengths of the formulas based on the interpolation algorithm used [2]. This paper focuses on explanations with convex impact space produced by interpolation algorithms based on Farkas' lemma [7] (Itp_F), its dual version (Itp'_F), and a flexible strength algorithm [8] (Itp_f) parametrized by a rational factor $f \in [0, 1]$ such that the logical strength of the yielded explanations follows $Itp_F \implies Itp_f \implies Itp'_F$.

Fig. 1 shows a 2D snapshot of an accurate explanation of a heart-attack risk prediction [9] based on the features age and cholesterol. Similarly, Fig. 2 demonstrates the capabilities of the framework to compute explanations with larger or smaller spaces based on different algorithms. However, they have limited means of comparison (e.g., visualization), as well as applicability for users to understand which feature correlations are important for drawing a medical conclusion.

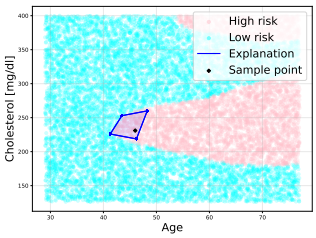

Fig. 1: Example of a space explanation (taken from [2])

We treat the impact space as a bounded closed *convex* set $\mathcal{X} \subset \mathbb{R}^m$ in Euclidean space, where m is a positive integer. This paper uses techniques from analytic geometry for the interpretation and comparison of logical explanations.

Maximum diameter [10] of $\mathcal{X}$ is a pair of points $(\boldsymbol{v}^i, \boldsymbol{v}^j)$ that are the solution of the problem $\arg\max_{\boldsymbol{v}^i, \boldsymbol{v}^j \in \mathcal{X}} \{||\boldsymbol{v}^i - \boldsymbol{v}^j||_2^2\}$. By Bauer's maximum principle [11], these points must be vertices of $\mathcal{X}$ since the squared norm is a convex function.

Minimum norm point is the solution of the problem $\arg\min_{\boldsymbol{x} \in \mathcal{X}} \{||\boldsymbol{x}||_2\}$. It is computed using quadratic programming optimization [12].

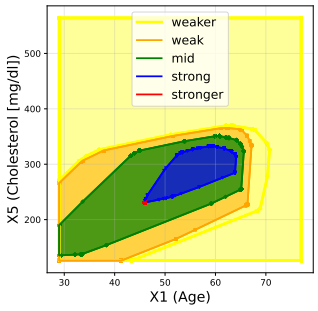

Fig. 2: Comparison of space explanations (from [2])

Hypervolume of $\mathcal{X}$ can be estimated using Monte Carlo method [13]. Uniform random point selection from a bounding space $\mathcal{X} \subseteq \mathcal{S} \subset \mathbb{R}^m$ implies that the probability of a point landing in $\mathcal{X}$ is proportional to the volume ratio $\frac{|\mathcal{X}|}{|\mathcal{S}|}$.

3 Interpretability and Comparison Methods

Extracting maximum diameter. Segments within the impact space of space explanations provide information about the underlying classifier: having a large diameter, such a direction indicates a linear relation between features along which the classification output remains locally unchanged. Algorithm 1 picks random pairs of points from the surface of $\mathcal{X}$ and maximizes over their Euclidean distance. It takes as input an explanation φ (with impact space $\mathbb{F}_\varphi$) and the number of random point choices N, and returns the best pair of surface points found within the budget. Function *select_on_surface* ensures each random point lies on the surface by projecting it onto a selected dimension. Deriving a linear relation among features is done by interpreting the coordinates of the resulting segment.

Algorithm 1: Maximum diameter estimation

Input: $\varphi, N, \mathbb{F}$ (with m features)
Output: $(\boldsymbol{y}^1, \boldsymbol{y}^2)$

real $best\_L \leftarrow 0$
list[real] $\boldsymbol{y}^1, \boldsymbol{y}^2$
for int $i \leftarrow 0$; $i < N$; $i \leftarrow i + 1$ **do**
  list[real] $\boldsymbol{x}^1 \leftarrow random.uniform(\mathbb{F})$
  $x^1_m \leftarrow select\_on\_surface(\boldsymbol{x}^1, \mathbb{F}_\varphi)$
  // analogously for $\boldsymbol{x}^2$ ...
  real $L \leftarrow ||\boldsymbol{x}^1 - \boldsymbol{x}^2||_2$
  if $L > best\_L$ **then**
    $best\_L \leftarrow L$
    $(\boldsymbol{y}^1, \boldsymbol{y}^2) \leftarrow (\boldsymbol{x}^1, \boldsymbol{x}^2)$
return $(\boldsymbol{y}^1, \boldsymbol{y}^2)$

However, if the variance of the length of all diameters is low, i.e., the impact space of the explanations resembles a spherical shape, the identified relation is not significant. To classify the shape of explanations, we compute the variance of the lengths using a set of standard uniformly distributed directions. A small variance (this paper uses 2%) classifies an explanation to be similar to a sphere.

Extracting minimum norm point. Minimum norm points serve as good representatives from which useful information can be extracted. In the context of image classification, such special points show how particular pixels participate in the classification process, while still accurately resembling the intended image. In particular, the minimum norm point allows to identify the subset of pixels that need to be greater than 0 for the classification.

Hypervolume estimation. Previous works [2, 1] compare the explanations based on the number of features and by checking whether one impact space is a subset of another. However, in many cases explanations are incomparable with those metrics. Hence, we propose a different metric using hypervolume estimation. It allows to reason about which explanation is more general and covers more points in the feature space. Consequently, it allows users to evaluate logic-based explanations without knowing logic. Algorithm 2 estimates the hypervolume of the impact space of the explanation using Monte Carlo method: given an explanation φ with impact space $\mathbb{F}_\varphi$ and the number of random points N, it computes the ratio of points that land inside of $\mathbb{F}_\varphi$. The ratio estimates the relative hypervolume of the impact space $\mathbb{F}_\varphi$ wrt. the size of the feature space $\mathbb{F}$.

Algorithm 2: Relative volume estimation

Input: φ (w. impact space $\mathbb{F}_\varphi$), N, $\mathbb{F}$
Output: $hypervolume \in [0, 1]$

int $h \leftarrow 0$
for int $i \leftarrow 0$; $i < N$; $i \leftarrow i + 1$ **do**
  list[real] $\boldsymbol{x} \leftarrow random.uniform(\mathbb{F})$
  if $\boldsymbol{x} \in \mathbb{F}_\varphi$ **then** $h \leftarrow h + 1$
return real $hypervolume \leftarrow (h/N)$

4 Experiments

The proposed methods were applied to the formal explainability study using a tabular dataset [9] that predicts heart attacks (low or high risk) based on 13 medical indicators of patients (a fully connected NN was trained using one hidden layer with 50 neurons and 85% accuracy) and MNIST dataset [14] that is a collection of grayscale handwritten digits (0–9) with 784 features (pixels) (a fully connected NN was trained using one hidden layer with 200 neurons and 96% accuracy). The tool SpEXplAIn[1] computed explanations for the classification of particular sample points: 303 points for the heart-attack task, and 50 points for the MNIST. Various interpolation algorithms[2] produced explanations of different logical strengths and sizes (from `strong` to `weak`) to provide multiple choices for the user. However, all of them are still machine-level complex and unreadable logical formulas. Our experimentation shows how meaningful information for the domain experts can be extracted even from such explanations.

The benefits of the extraction of the *maximum diameter* from logical explanations are evident in the medical domain task. Fig. 3 shows how the maximum diameters representation of the explanation is easier to interpret. Each subfigure represents a space explanation (in purple) in the area around the decision boundary between the low-risk (blue) and the high-risk (orange) class spaces, and the maximum diameter (red) for selected pairs of input features of the NN (e.g., age and cholesterol)[3]. The domains are normalized. The experimentation has been done using explanations of different strengths, arriving at similar observations that the extracted diameter is indeed a meaningful metric to characterize space explanations; the figure shows `weak` explanations that are most spacious and illustrative. Each maximum diameter reveals a linear relation between the input features of the NN and can guide the user to draw further conclusions. Notably, while being understandable to the user, they remain in the form of logical formulas that guarantee the classification. For example, the logical formula describing the maximum diameter in Fig. 3a is $\varphi_{\mathrm{a}} = (0.3 \leq X1 \leq 1) \wedge (X5 = -0.096X1 + 0.24)$. These formulae are way less verbose than the original explanations. However, the relation may be insignificant in cases when the explanation resembles a spherical shape (e.g., Fig. 3e and f). To classify such cases, we computed the variance of the length of diameters. For our illustrative example, explanations e–h are classified as spherical. Notably, the runtime overhead of this computation is negligible.

The *extraction of minimum points* serves for interpreting and comparing logical explanations visually. For example, in the digit classification task MNIST, we computed the minimum norm points that are suitable for identifying relevant pixels because the background is black (if the colors were inverted, the appropriate choice would be the maximum norm point). Fig. 4 illustrates the relevance of particular pixels in explanations of different strengths: `weak` in the

[1] `https://github.com/usi-verification-and-security/spexplain`

[2] The interpolation algorithms use the notation $Itp_F \mapsto$ `strong`, $Itp'_F \mapsto$ `weak`, and $Itp_f \mapsto$ `mid` with $f := 0.5$.

[3] The figures show 2D snapshots for clarity, but the approach is general for more dimensions.

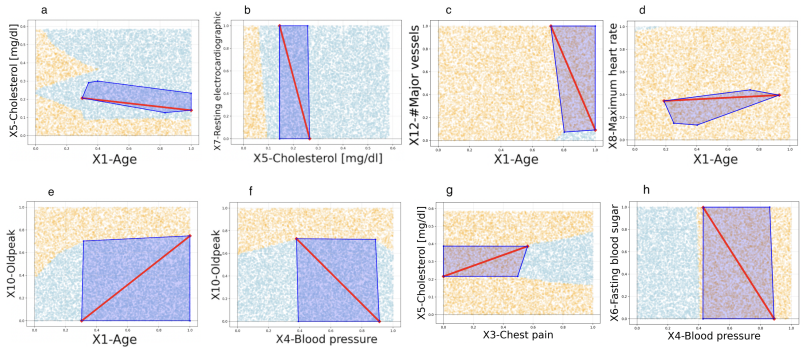

Fig. 3: Maximum diameters (red) of space explanations (purple)

first row, and `strong` in the second row. The first row results in a sparse and fuzzy distribution of pixels, while the second row results in a dense distribution, suggesting that those explanations are more suitable. Notably, the norm points are not just visualizations but still logical explanations guaranteeing the classification—in contrast to, e.g., saliency maps [15].

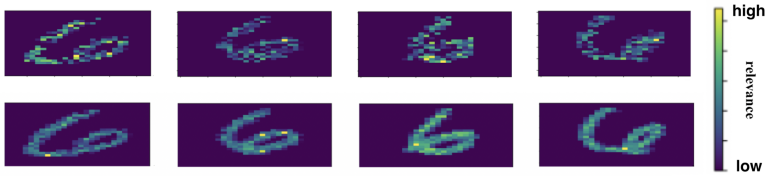

Fig. 4: Visualization of minimum norm points within space explanations

Volume	Full	2D
`strong`	0.0000077	0
`mid`	0.0003440	0.34
`weak`	0.0035000	0.49

Table 1: Average relative hypervolume

	Full-D		2D	
# Points	Error	Time	Error	Time
10^5	15%	3.6s	3%	2.3s
10^6	7%	10.2s	0.5%	6.29s
10^7	1%	97.3s	0.02%	51.23s

Table 2: Error and runtime of volume estimation

The *hypervolume estimation* was applied to the medical domain. Table 1 illustrates the results of the comparison of the average relative hypervolume for all explanations of given logical strength. The second column shows the volumes of the entire explanations. The third shows the average over the 2-dimensional snapshots of all pairs of input features. The hypervolume of stronger explanations is, as expected, significantly lower than the hypervolume of weaker explanations. In many cases, the hypervolume of `strong` explanations is zero because some of the features are fixed to a single value (also reported in [2, Table 1]). Notably, the proposed comparisons are possible even in cases where the explanations do not originate from the same sample points or do not even

intersect. Table 2 illustrates the dependency between the mean error of the estimated hypervolume and the runtime using `mid` explanations (the numbers are similar for the others). As expected, the error decreases with more random points, and the runtime increases linearly.

5 Conclusions

The paper addressed a problem of interpreting the explanations produced by the application of logical techniques to machine learning models. It proposed the use of classical geometric methods to make the logical explanations human friendly and usable. The advantages of the proposed methods were illustrated on the benchmarks already thoroughly studied by the formal verification community which, while considered the state-of-the-art, remained hard to comprehend by the users of the NNs. Remarkably, the geometric interpretations of the logical explanations do not incur significant computational overhead and thus can be integrated with the formal verification explainability naturally.

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

References

[1] M. Wu, H. Wu, and C. Barrett. VeriX: Towards Verified Explainability of Deep Neural Networks. In *NeurIPS*, 2023.

[2] F. Labbaf, T. Kolárik, M. Blicha, G. Fedyukovich, M. Wand, and N. Sharygina. Space explanations of neural network classification. In *CAV*, 2025.

[3] Y. Izza, A. Ignatiev, P. J. Stuckey, and J. Marques-Silva. Delivering Inflated Explanations. In *EAAI*, 2024.

[4] A. Ignatiev, N. Narodytska, and J. Marques-Silva. Abduction-Based Explanations for Machine Learning Models. In *EAAI*, 2019.

[5] M. Wu, X. Li, H. Wu, and C. Barrett. Better verified explanations with applications to incorrectness and out-of-distribution detection. In *arXiv*, 2024.

[6] W. Craig. Three Uses of the Herbrand-Gentzen Theorem in Relating Model Theory and Proof Theory. In *J. of Symbolic Logic*, 1957.

[7] M. Blicha, A. Hyvärinen, J. Kofron, and N. Sharygina. Decomposing Farkas interpolants. In *TACAS*, 2019.

[8] L. Alt, A. Hyvärinen, and N. Sharygina. LRA interpolants from no man's land. In *HVC*, 2017.

[9] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano. Heart Disease, 1989.

[10] B. Grunbaum. *Convex Polytopes*. Springer, 2003.

[11] H. Bauer. Minimalstellen von Funktionen und Extremalpunkte. *Archiv der Mathematik*, 1958.

[12] J. Nocedal and S. J. Wright. *Quadratic Programming*. Springer New York, 2006.

[13] N. Metropolis and S. Ulam. The monte carlo method. *Journal of the American Statistical Association*, 1949.

[14] L. Deng. The MNIST Database of Handwritten Digit Images for Machine Learning Research. In *EAAI*, 2012.

[15] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In *ICLR*, 2014.