

Graph theory based mathematical methods for improved explainable AI

Ansh Vashist [@]

Aditya Joshi [†]

Department of Computer Science Engineering

Manipal University Jaipur

Jaipur

Rajasthan

India

Sanjay Tiwari [§]

Department of Computer Science Engineering

Arya Institute of Engineering Technology and Management

Jaipur

Rajasthan

India

Rakesh Sharma [†]

Department of Computer Science Engineering

Vivekanand Global University

Jaipur

Rajasthan

India

Tarun Jain ^{*}

Department of Computer Science Engineering

Manipal University Jaipur

Rajasthan

India

[@] E-mail: anshvst2907203@gmail.com

[†] E-mail: ajoshi2002@gmail.com

[§] E-mail: sanjay76tiwari@gmail.com

[‡] E-mail: srak.911@gmail.com

^{*} E-mail: tarunjainjain02@gmail.com (Corresponding Author)

Pankaj Dadheech [#]

Department of Computer Science and Engineering

Swami Keshvanand Institute of Technology, Management and Gramothan

Jaipur

Rajasthan

India

Abstract

This paper introduces a novel framework aimed at improving the clarity and dependability of explanations generated by AI systems specializing in Natural Language Processing (NLP). By employing a comprehensive graph-theoretical and mathematical methodology, the proposed model systematically enhances both the accuracy and reliability of AI's explanatory functions. Central to this methodology is the Accuracy-Reliability Enhancement Framework (AREF), which utilizes graph theory to represent and analyze the AI decision-making process. This approach enables the identification and refinement of explanation pathways that maintain high fidelity and consistent reliability across various inputs. A specific objective function is developed to meticulously balance the need for high explanation fidelity and consistency against the complexity of the explanatory model. The framework leverages advanced mathematical capabilities of graph theory, including path analysis and optimization, node importance evaluation, and key subgraph identification. These features ensure that the explanations are not only transparent and easy to understand but also based on solid mathematical principles. By integrating graph-theoretical and mathematical approaches with the goals of Explainable AI (XAI), this approach significantly advances the development of AI systems that are both comprehensible and trustworthy, addressing a crucial need in contemporary XAI techniques through a clearly defined, mathematically rigorous strategy.

Subject Classification: Primary 65A05, Secondary 26C10.

Keywords: *Explainable AI, Natural language processing, Graph theory decision, Interpretability in AI, Fidelity-consistency optimization.*

1. Introduction

The quest for transparency and understandability in Artificial Intelligence (AI), particularly in Natural Language Processing (NLP) systems, has led to a burgeoning interest in Explainable AI (XAI). As NLP systems become increasingly integral to various sectors, including healthcare, finance, and legal, the imperative for these systems to not only make decisions but also provide comprehensible explanations for their actions has become paramount. However, the complexity inherent in

[#] E-mail: pankajdadheech777@gmail.com

modern AI models, especially those employing deep learning techniques, poses a significant challenge to this endeavor. These models, often regarded as “black boxes,” offer limited insight into their decision-making processes, thereby impeding trust and reliability [1]. To address this challenge, our work introduces a novel theoretical framework designed to enhance the interpretability and reliability of AI explanations. At the heart of our proposition is the Fidelity-Consistency Optimization Framework (FCOF), which leverages the mathematical rigor of graph theory to articulate and optimize the pathways through which AI systems arrive at their conclusions. Graph theory, with its rich vocabulary for describing complex networks of relationships and interactions, provides a powerful tool for modeling the decision-making processes of AI systems. By conceptualizing these processes as graph-based pathways, our approach facilitates the generation of explanations that are both faithful to the AI’s operations (fidelity) and consistent across similar or slightly varied inputs (consistency).

In the rapidly evolving field of artificial intelligence (AI), the imperative for explainable AI (XAI) has become increasingly paramount. Explainability not only fosters trust and transparency but also ensures regulatory compliance and facilitates debugging and model improvement. Central to the utility of XAI is the optimization of two critical dimensions: fidelity, which measures the accuracy of explanations in reflecting a model’s decision-making process; and consistency, which assesses the stability of explanations across similar instances [2]. The strategies and mechanisms for optimizing these dimensions, offering a novel perspective on balancing the trade-offs as objective of this research.

The mathematical underpinnings of our framework employ graph-theoretic concepts such as path enumeration, which identifies all possible routes through a decision graph; node centrality, which assesses the importance of various components in the decision-making process; and subgraph extraction, which isolates the most relevant portions of the graph for generating explanations. These features enable our model to systematically identify the optimal balance between fidelity, consistency, and the simplicity of explanations, thereby addressing one of the most pressing challenges in XAI: the development of interpretable and reliable explanatory mechanisms that are accessible to end-users without sacrificing the model’s complexity or performance. Incorporating mathematical expressions, our framework might define a decision graph $G=(V,E)$, where V represents the set of nodes (decision points, features, or outcomes) and E represents the edges (decision paths or relationships)

between features). The fidelity of an explanation can be quantified using a function $f:V \rightarrow R$, mapping each node to a fidelity score, while consistency can be modeled as a constraint optimization problem, seeking to minimize the variance in explanation paths for similar inputs, formalized as $\min \text{Var}(E)$, where $\text{Var}(E)$ represents the variance in the edges selected for explanations across different but similar inputs. In deploying this graph-theoretic approach, our work not only contributes a novel methodology to the field of XAI but also establishes a foundation for further research into the development of transparent, understandable, and mathematically grounded AI systems. This introduction sets the stage for a detailed exploration of the FCOF, its mathematical features, and its potential to revolutionize the way we understand and interact with AI systems in NLP and beyond.

1.1 Theoretical background

Fidelity and consistency serve as fundamental benchmarks in evaluating the effectiveness of XAI techniques. High fidelity ensures that the explanations accurately represent the model's behavior, while high consistency guarantees that the explanations are robust across different instances or slight variations in input data. Achieving an optimal balance between these two dimensions is challenging due to their often-competing nature; enhancing one aspect can sometimes come at the expense of the other [3].

The optimization trajectory within this graph illustrates the pathway from initial states of low fidelity and consistency towards a goal state where both are maximized. This trajectory highlights the iterative process of applying different strategies, such as the incorporation of more sophisticated local explanation models, the use of regularization techniques to enhance stability, and the iterative refinement of explanations based on feedback loops [4].

2. Evaluating XAI through Fidelity and Consistency

Explainable Artificial Intelligence (XAI) seeks to make the operations of AI systems transparent and understandable to human users. This transparency is crucial for building trust, facilitating user decision-making, and ensuring the ethical deployment of AI technologies. Within this context, fidelity and consistency emerge as fundamental benchmarks for

evaluating the effectiveness of XAI techniques [5]. Their roles in shaping the reliability and usefulness of explanations cannot be overstated.

2.1 *Fidelity: Accurate Interpretations*

Fidelity refers to the degree to which the explanations generated by an XAI technique accurately reflect the actual workings and decisions of the underlying AI model. High fidelity ensures that the explanations are truthful representations, providing insights into the model's decision process [6]. This is paramount for users who rely on these explanations to understand model behavior, diagnose potential issues, or justify decisions made by AI systems in critical applications.

Mathematically, fidelity can be quantified as the similarity between the output of the AI model and the explanation model for the same input. Let $f(x)$ be the output of the AI model for input x and let $g(x)$ be the corresponding explanation. Fidelity F can be expressed as a measure of similarity S between $f(x)$ and $g(x)$:

$$F = S(f(x), g(x)) \quad \dots(i)$$

where higher values of F indicate better fidelity.

2.2 *Consistency: Ensuring Reliability*

Consistency measures the stability or uniformity of explanations across similar instances or under slight variations in input. In the context of XAI, it ensures that the explanations provided for similar decisions by the AI model are coherent and do not contradict each other. This is crucial for user trust, as inconsistent explanations can lead to confusion and undermine confidence in the AI system's reliability.

Consistency can be mathematically defined for a pair of similar inputs x_1 and x_2 , leading to explanations $g(x_1)$ and $g(x_2)$, respectively. Let D denote a distance metric that measures the dissimilarity between explanations. Consistency C can then be inversely related to the average distance between explanations for similar inputs:

$$C = 1/D(g(x_1), g(x_2)) \quad \dots(ii)$$

where higher values of C reflect better consistency, indicating that explanations for similar inputs are more alike.

The effectiveness of XAI techniques hinges on achieving a delicate balance between fidelity and consistency. High fidelity without consistency may lead to accurate but unpredictable explanations, while consistent but low-fidelity explanations might fail to accurately represent the AI model's

decision-making process. Optimal XAI techniques strive to maximize both, providing explanations that are both accurate and reliably stable across similar cases.

3. Strategies for Optimization

In the quest for optimizing explainable AI (XAI), the strategies for optimization play a pivotal role in enhancing the interpretability, reliability, and overall effectiveness of AI systems. These strategies, including the implementation of local explanation models, the application of regularization techniques, and the incorporation of feedback loops, are designed to address the dual objectives of maximizing fidelity and consistency in explanations [7]. Local explanation models focus on approximating the decision-making process of AI within a localized context, offering precise and tailored insights. Regularization techniques aim to stabilize the explanations across similar instances, ensuring that the interpretability of AI systems remains consistent and reliable. Feedback loops engage human experts in the loop, allowing for the continuous refinement of explanations based on real-world insights and perceptions. Collectively, these strategies form a comprehensive approach to overcoming the challenges inherent in explaining complex AI models, paving the way for more transparent, trustworthy, and user-friendly AI technologies [8].

3.1 Local Explanation Models

Adopting local explanation models that approximate the global model's decision boundary can improve fidelity without substantially compromising consistency [9]. Local explanation models aim to approximate the decision boundary of a complex global model within a local vicinity of a given instance. Let $f(x)$ represent the global model's prediction for an input x , and let $g(x')$ denote the local model's prediction for a perturbed input ' x' ' near x . The goal is to minimize the difference between $f(x)$ and $g(x')$ for ' x' ' about x , which can be formalized as:

$$\text{Min}_{g \in G} \sum_{x' \in N(x)} L(f(x'), g(x')) + \Omega(g) \quad \dots(\text{iii})$$

where L is a loss function measuring the discrepancy between the predictions of the global model and the local model on the neighborhood $N(x)$ of x , G is the space of possible local models, and $\Omega(g)$ is a regularization term for the complexity of the local model g .

3.2 Regularization Techniques

Implementing regularization techniques in the explanation generation process can enhance consistency by penalizing large deviations in explanations for similar instances. Regularization techniques aim to enhance the stability (consistency) of explanations across similar instances by penalizing large deviations in the explanations [10]. For an explanation model E generating an explanation e for an input x , and a similar input ' x ' with explanation ' e ', the regularization can be introduced as a penalty term in the optimization objective:

$$\text{Min}_E L(E(x), y) + \lambda \cdot R(e, e') \quad \dots(\text{iv})$$

where L is the loss function measuring the fidelity of the explanation $E(x)$ to the true output y , $R(e, e')$ is a regularization term measuring the dissimilarity between explanations for similar inputs, and λ is a coefficient balancing the trade-off between fidelity and consistency.

4. Proposed Graph-Theoretic Framework

In a graph-theoretic framework, the landscape of possible explanations is conceptualized as a graph, where nodes represent distinct states of the explanation model, characterized by specific levels of fidelity and consistency. Edges between these nodes symbolize transitions that either enhance or compromise these attributes. Such a framework facilitates the systematic exploration of the explanation space, guiding the selection of optimal paths that lead to high-quality explanations.

Let $G = (V, E)$ denote the graph, where V is the set of nodes (states) and E is the set of edges (transitions). Each node $v \in V$ is associated with a tuple $(F(v), C(v))$, representing the fidelity and consistency of the explanation at that state. The goal is to navigate from an initial state v_0 to a target state v^* that maximizes both $F(v^*)$ and $C(v^*)$.

The optimization process involves identifying a path P from v_0 to v^* that optimizes a combined objective function $\Phi(F, C)$, reflecting a balance between fidelity and consistency. This path represents a sequence of transitions (edges) that progressively enhance the quality of explanations.

$$\text{Max} P = (v_0, \dots, v^*) \Phi(F(v^*), C(v^*)) \quad \dots(\text{viii})$$

Graphical Illustration

Nodes represent states of explanations, each with varying levels of fidelity and consistency.

Edges indicate possible transitions between these states, influenced by applying different XAI techniques or adjustments.

The figure 1 visually represents the concept of navigating through a space of explanations within a graph-theoretic framework to optimize fidelity and consistency in explainable AI (XAI). Each node A to E symbolizes a state of the explanation with varying levels of fidelity and consistency, where the goal is to move from an initial state with lower fidelity and consistency (e.g., node A) towards a state with higher values (e.g., node E).

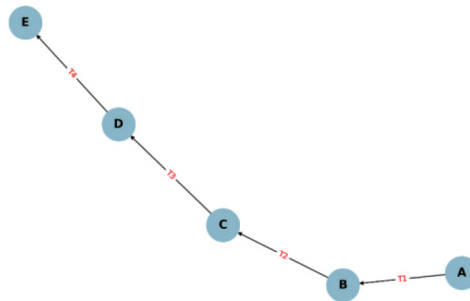


Figure 1
Varying levels of fidelity and consistency

Nodes: Indicate states of explanation quality, with attributes for fidelity (how accurately the explanation reflects the model's operation) and consistency (how stable the explanation is across similar inputs).

Edges (T1 to T4): Represent transitions or actions that can be taken to improve the explanation's fidelity and consistency. These transitions could involve refining the explanation model, adjusting the parameters, or incorporating additional information to enhance interpretability.

This graph-theoretic approach enables the systematic exploration of the explanation space, allowing for the identification of optimal paths that lead to the most informative, accurate, and reliable explanations. Through this framework, we can quantitatively measure and incrementally improve the quality of explanations, thereby contributing significantly to the advancement of explainable AI.

Conclusion

The proposed graph-theoretic framework represents a significant advancement in the pursuit of optimizing fidelity and consistency within explainable AI (XAI). By conceptualizing the space of explanations as a

graph, where states denote various levels of explanation quality and edges embody potential improvements, this approach provides a structured methodology for navigating towards the most effective and reliable explanations. It not only facilitates a deeper understanding of the intricate balance between fidelity and consistency but also offers a practical roadmap for enhancing the transparency and trustworthiness of AI systems. Through the implementation of this framework, we envision a future where AI's decisions are not only more interpretable and accessible to users but are also aligned with the highest standards of accuracy and consistency. This work lays the groundwork for further research and development in XAI, opening new avenues for innovation and the ethical deployment of artificial intelligence technologies.

References

- [1] M. Li, Y. Zhan, and Y. Ge, "Fractional distance regularized level set evolution with its application to image segmentation," *IEEE Access*, vol. 8, pp. 120–130 (2020).
- [2] J. R. Clough, N. Byrne, I. Oksuz, V. A. Zimmer, J. A. Schnabel, and A. P. King, "A topological loss function for deep-learning based image segmentation using persistent homology," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8766–8778, Dec. (2020).
- [3] A. Runacher, M. J. Kazemzadeh-Parsi, D. Di Lorenzo, V. Champaney, N. Hascoet, A. Ammar, and F. Chinesta, "Describing and modeling rough composite surfaces by using topological data analysis and fractional Brownian motion," *Polymers*, vol. 15, no. 6, pp. 1449 (2023).
- [4] S. Lee, H. Kim, Q. X. Lieu, and J. Lee, "CNN-based image recognition for topology optimization," *Knowl.-Based Syst.*, vol. 198, pp. 105887 (2020).
- [5] D. Youstri, M. A. Elaziz, and S. Mirjalili, "Fractional-order calculus-based flower pollination algorithm with local search for global optimization and image segmentation," *Knowl.-Based Syst.*, vol. 197, pp. 992–1001 (2020).
- [6] C. Biswas, T. Biswas, and R. Biswas, "Relative (p, q) - ϕ order and relative (p, q) - ϕ type oriented some growth investigations of composite p -adic entire functions," *J. Interdiscip. Math.*, vol. 25, no. 2, pp. 233–246 (2022).

- [7] E. Courtois, *Topological Data Analysis of Convolutional Neural Networks Using Depthwise Separable Convolutions*, Ph.D. dissertation, University of Missouri-Saint Louis (2023).
- [8] M. Nishio, M. Nishio, N. Jimbo, and K. Nakane, "Homology-based image processing for automatic classification of histopathological images of lung tissue," *Cancers*, vol. 13, no. 6, pp. 1192 (2021).
- [9] Slimane, Ibrahim, and Zoubir Dahmani. "Normalized fractional inequalities for continuous random variables," *Journal of Interdisciplinary Mathematics*, Vol. 25.2, pp. 335-349 (2022).
- [10] J. R. Clough, N. Byrne, I. Oksuz, V. A. Zimmer, J. A. Schnabel, and A. P. King, "A topological loss function for deep-learning based image segmentation using persistent homology," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8766–8778, Dec. (2020).

Received February, 2024