Exposing Inductive Biases of Deep Graph Networks through Explainable AI

Michele Fontanesi[1,*,†], Alessio Micheli[1,†] and Marco Podda[1,†]

[1]University of Pisa, Department of Computer Science, Largo B. Pontecorvo 3, 56127 Pisa, Italy

Abstract

The field of Explainable Artificial Intelligence (XAI) for Deep Graph Networks (DGNs) collects methods to study the learned correlation between the input graphs and their labels. The extracted information is then provided as an explanation to increase the user's trust in the system's response. However, the purpose of these techniques extends beyond the search for explanations. In this short abstract, we provide an overview of some research directions that stem from the field of XAI for DGNs, contextualizing their relevance for the fields of XAI and DGN and their pertinence to the Ph.D program. Then, we provide further details on the main concepts behind a methodological approach, based on XAI techniques, to study the inductive biases of diverse DGN variants performing graph classification tasks while offering a synopsis of the acquired findings.

Keywords

Deep Graph Networks, Explainable AI, Inductive Biases

1. Introduction

Graphs are complex data structures, comprising entities, or vertices, associated pairwise through relationships that are modeled as edges. As vertices and edges may assume any type of semantic, graphs are a very flexible modeling approach but their non-euclidean structure makes them hard to process and study. Deep Graph Networks [1], pioneered by [2] and [3], are currently the most powerful, versatile, and promising approach to solve classification as well as regression tasks on graph data [4, 5, 6]. However, DGNs are still far from being a human-centered approach as the logic behind their responses is hidden in the learned parameters. This lack of transparency hinders their trustworthiness [7], and consequently, their adoption, as understanding an autonomous system's response became imperative [8]. To this end, the field of Explainable AI (XAI) [9] has been founded and a huge research effort has been put into developing techniques able to highlight the correlation learned by Neural Network, including DGNs [10], between the input data and the target labels. The retrieved information is then mainly used to craft an explanation regarding the reasons behind the model outcome. However, the importance of XAI techniques for DGNs overcomes the sole objective of explaining to the final user, as it is possible to identify more purposes for these methods and, consequently, diverse research directions across the fields of XAI and DGNs. These directions are all aligned with the objectives of the National Ph.D. in Artificial Intelligence for Society as any advancements in the fields of XAI or DGN would either close the gap between humans and AI or provide humans with better tools to address and understand complex problems. We outline these research directions in Section 2, highlighting their relevance and impacts in the fields of XAI and DGNs. Further details are provided for the research direction that is currently under active investigation. For this latter one, the background to the methodological approach is outlined in section 3, while the methodology itself is summarised in section 4. In section 5, we introduce the preliminary results, while in section 6 we discuss future research activities.

Doctoral Consortium at the 23rd International Conference of the Italian Association for Artificial Intelligence Bolzano, Italy, November 25-28, 2024.

*Corresponding author.

†These authors contributed equally.

michele.fontanesi@phd.unipi.it (M. Fontanesi); alessio.micheli@unipi.it (A. Micheli); marco.podda@unipi.it (M. Podda)

0009-0004-7566-903X (M. Fontanesi); 0000-0001-5764-5238 (A. Micheli); 0000-0003-1497-9515 (M. Podda)

2. Reserach Directions

XAI for knowledge extraction. A DGN able to successfully solve a task can be seen as a possible source of information to derive new knowledge concerning the problem it has learned to solve [6]. To this matter, multiple and different XAI techniques can be applied with the purpose of retrieving the meaningful patterns that a DGN has learned to associate with a particular class [11]. Among the patterns, some may reveal novel insights into the faced problem. This line of work addresses the following research questions: are different XAI techniques converging to the same or similar explanations? Are retrieved explanations meaningful to acquire new knowledge?

XAI-based architectures. Most XAI techniques for DGNs are model-independent approaches capable of analyzing a trained DGN at test time (post-hoc) [12, 13]. However, an interesting, promising, and challenging research direction is to find architectural design principles that make DGNs easier to analyze for different tasks and at different levels of granularity: input-output-wise, layer-wise, and unit-wise. A direct consequence of such a design is the creation of a direct coupling between the architecture and the extracted information which increases the reliability of the analysis and therefore the trust in the retrieved explanation. This line of work addresses the following research question: can we retrieve more meaningful explanations by introducing the explainability requirement into the design of a DGN approach?

XAI for model analysis and improvement. Post-hoc, model-agnostic techniques [12, 13] may be used to study the behavior of different DGN models on a given task and to identify differences and potential shortcomings of each architecture. The acquired knowledge could be exploited to understand which models are better at solving a given task and to learn why some architectures achieve better performances. This line of work addresses the following research questions: are feedforward [14], constructive [2], and recursive [15] architectures solving tasks based on the same input patterns? Can we understand which DGN variants are better suited to solve a task based on the input graph properties?

This research direction stems from the observation that XAI techniques can be used as model inspection tools to analyze the diverse inductive biases characterizing different types of DGNs. Inductive biases are the set of assumptions used by a DGN to perform predictions on unknown inputs and consequently, their characterization is of the utmost importance to select the model that better aligns with a particular learning task to solve. In this regard, we have demonstrated that XAI methodologies can be utilized to discern the class assignment policy induced by the inductive biases and learned by each type of DGN to associate a graph with its target class. Specifically, we investigated the inductive biases of recursive and convolutional DGNs in graph classification tasks. This was achieved by comparing the explanations generated by XAI techniques with the ground truth (GT) explanations associated with each valid policy. Results highlighted (i) the existence of diverse class assignment policies for three XAI graph classification benchmarks [16, 17] [18], (ii) the capabilities of recursive and convolutional DGNs to learn different policies [18], (iii) the effect of using multiple layers on the inductive bias of convolutional DGNs [19], and (iv) the alignment of recursive and convolutional DGN explanations with the values of Katz centrality and Fiedler eigenvector, respectively [20]. From the XAI side, characterizing the inductive biases of diverse DGNs may increase the trust in these systems as we identify the problem specifics that diverse DGNs can leverage to solve a task. From an ML point of view, linking the aspects learned by diverse DGN variants with their specific formulation may be beneficial to developing more performant and efficient models.

3. Background

Deep Graph Networks. A DGN is a parameterized function capable of learning a mapping between input graphs $G \in \mathcal{G}$ and their associated classes $y \in \mathcal{C}$ following the message passing (MP) paradigm; a procedure that updates node embeddings $\mathbf{h}_v \in \mathbb{R}^d$ (the vectorial information associated with each node)

iteratively starting from the initial node feature vectors $\mathbf{x}_v \in \mathbb{R}^k$. MP is a blueprint defined at the node level as follows:

$$\mathbf{h}_v^{l+1} = \text{Upd}\left(\mathbf{h}_v^l,\ \text{Agg}(\{\text{Msg}(\mathbf{h}_v^l,\ \mathbf{h}_u^l) \mid u \in \mathcal{N}_v\})\right), \tag{1}$$

where the Msg function computes a message between every node and its neighbors; Agg summarizes all the messages received by each node in a permutation-invariant fashion; and Upd combines the current node embedding and the aggregated messages to generate a novel embedding for each node. DGN characteristics are determined by their specific implementation of the MP blueprint. Across the set of experiments, we studied convolutional DGN variants as GIN (Graph Isomorphism Networks) [21], GC (Graph Conv) [22] and PNA (Principal Neighborhood Aggregation) [23] and, as recursive variant, GESN (Graph Echo State Networks) [15]. Each variant features a pooling operator to generate a single graph vector based on which each model outputs the target class probabilities.

XAI attribution methods. Across the experiments summarized in this short abstract, we used a local post-hoc XAI technique able to associate an *importance score* with each node in a graph in the form of a mask $\hat{\mathbf{m}} \in \mathbb{R}^{N_G}$ with N_G the cardinality of the set of nodes of graph G. Among the many methods that exist in the literature [12, 13], we employed CAM [24] as we observed that it was able to compute more stable explanations than GNNExplainer [10] or Integrated Gradients [25].

Graph centrality and connectivity notions. The Katz centrality [26] values are higher for the nodes that have in their neighborhood many other well-connected nodes. As a consequence, the notion of Katz centrality is well suited to detect an inductive bias that leads DGNs to solve the graph classification tasks based on low-order graph structures like isolated nodes with a high degree. However, a DGN may also base its predictions on higher-order structural information as detecting the presence of certain subgraphs. To Identify this second type of inductive bias through node scores we used the values of the the Fiedler eigenvector [27] whose signs are usually used to cut the graph into two communities.

4. Method

Our methodology is based on multiple XAI graph classification datasets of the form $\mathcal{D} = \{(G, y_G, \mathcal{T}) \mid G \in \mathcal{G}, y \in \mathcal{C}\}$ where graphs G are associated to target classes y as well as to sets of *ground truth explanations* $\mathcal{T} = \{\mathbf{m}_G^p \in \{0,1\}^n \mid p \in \mathcal{P}\}$ collecting a diverse *ground truth* (GT) for each class assignment policy in the set $\mathcal{P}$. In particular, a GT explanation is a binary vector that encodes the relevance (1) or irrelevance (0) of each node to the graph prediction depending on the associated class assignment policy p. To identify the policy learned from each DGN variant (trained with cross-validation), we computed the explanations for each test sample and quantified their adherence to the available GT with the plausibility score [28] (AUROC). Then, we identified the policy learned by a DGN as the one associated with the GT that maximizes the average plausibility scores across the test set samples. Last, we compute the average Pearson Correlation Coefficient between the explanation importance scores and the Katz centrality and Fiedler values to identify whether the inductive biases of diverse DGNs focus on low or high-order graph structure, respectively.

5. Preliminary Results

First, we discovered the existence of diverse class assignment policies for the XAI graph classification datasets of BA2Motif [17], BA2grid [16] and GridHouse [16]. In particular, we found that the correct graph class could be predicted by either looking at the presence of a motif or by identifying the nodes with a degree greater or equal to three. In figure 1 we provide as an example the GT explanations associated with the motif-based assignment policy and the degree-based assignment policy.

Then, through the computation of the plausibility metric, we found that GESN and PNA are characterized by a strong inductive bias that leads them to always learn the degree-based policy. GC and

(a) Motif-based (b) Degree-based

Figure 1: Example of two possible GT explanations. In both cases, the graph is assigned class 1 but in (a) this is done by retrieving the motif, while in (b) this is done by focusing on the nodes with degree ≥ 3.

GIN, instead, were capable of learning a different policy depending on the number of layers of their architecture and the local minima reached by the optimization procedure, as shown in Figure 2.

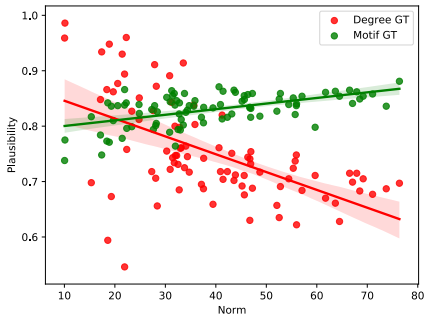

Figure 2: Average plausibility trends of the degree-based and the motif-based GT with respect to the 2-norm of the learned weights for a 2-layer GIN architecture on the BA2Motif dataset.

Last, we computed the average Pearson correlation coefficients between the explanation scores of various types of DGNs and the Katz centrality and Fiedler values. The obtained results highlighted the better alignment between the Katz centrality and Fiedler values with the explanation scores computed for the recursive and convolutional DGN variants, respectively.

6. Discussion and future research activities

In this short abstract, we introduced some research directions related to the fields of XAI and DGNs while focusing on the one that is currently under investigation. For this latter research direction, we summarized our contributions to the DGNs and XAI fields. In particular, we found that (i) simple graph classification tasks can feature multiple class assignment policies as viable solutions, (ii) recursive and convolutional DGNs feature diverse inductive biases that lead them to learn a preferred class assignment policy, (iii) the learned policy is influenced by the architectural number of layers and, in some cases, by the training procedure and, (iv) that inductive biases may be grounded in known concepts of graph theory as the Katz centrality and the Fiedler values. Studying and characterizing the inductive biases of DGNs impacts both the fields of XAI and DGNs. From the XAI perspective, increasing the knowledge about the inductive bias and consequently, the generalization capabilities of different DGNs variants leads to a more conscious and trustful application of these methods to different tasks. Moreover, the discovery of multiple sound GT raises warnings on the benchmarking processes of the XAI attribution methods as lower performance may be due to the usage of the wrong GT. From the DGN perspective, instead, understanding the association between the inductive biases and the MP variants may uncover opportunities to create novel models. In addition, from a practical

perspective, our results may be of use to practitioners in selecting the DGN variant that best aligns with the characteristics of the task they want to solve. As future research directions, we plan to perform extensive experiments by (i) increasing the number of tested DGNs including spectral [29], constructive [2], and other convolutional variants [30], (ii) increasing the number of tested explainers including generative [31] and factual/counterfactual approaches [32], and (iii) increasing the number of tested datasets possibly featuring non-synthetic graphs. In particular, extending results to more DGN variants would facilitate the discovery and characterization of additional opportunities to solve and generalize on graph-related tasks. Increasing the number of explainers would help to explore and compare different explanations. Finally, adopting real-world datasets would help in understanding the DGNs and explainer's behaviors outside controlled synthetic environments. However, the required datasets should feature ground truth explanations to check whether a DGN coupled with a particular explainer was aligned with the problem characteristics. We plan to find graphs with associated GT by exploiting the knowledge already developed in bioinformatics and chemistry. Alternatively, the field of business optimization processes can provide graphs modeling concurrent and interacting procedures with GT retrieved from the knowledge developed in the field. We also expect that achievements along this research direction may become opportunities to start investigating the direction of "XAI for knowledge extraction" in the fields of bioinformatics and chemistry and the direction of "XAI-based architecture" by exploiting the knowledge acquired on the tested explainers and inductive biases of DGNs.

Acknowledgments

Research partly funded by PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI", funded by the European Commission under the NextGeneration EU programme.

References

[1] D. Bacciu, F. Errica, A. Micheli, M. Podda, A gentle introduction to deep learning for graphs, Neural Networks 129 (2020) 203–221. URL: https://www.sciencedirect.com/science/article/pii/S0893608020302197. doi:https://doi.org/10.1016/j.neunet.2020.06.006.

[2] A. Micheli, Neural network for graphs: A contextual constructive approach, IEEE Transactions on Neural Networks 20 (2009) 498–511. doi:10.1109/TNN.2008.2010350.

[3] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, G. Monfardini, The graph neural network model, IEEE Transactions on Neural Networks 20 (2009) 61–80. doi:10.1109/TNN.2008.2005605.

[4] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in: Proceedings of the AAAI conference on artificial intelligence, volume 32, 2018.

[5] A. Derrow-Pinion, J. She, D. Wong, O. Lange, T. Hester, L. Perez, M. Nunkesser, S. Lee, X. Guo, B. Wiltshire, et al., Eta prediction with graph neural networks in google maps, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 2021.

[6] M. Fontanesi, A. Micheli, P. Milazzo, M. Podda, Exploiting the structure of biochemical pathways to investigate dynamical properties with neural networks for graphs, Bioinformatics 39 (2023) btad678. doi:10.1093/bioinformatics/btad678.

[7] L. Oneto, N. Navarin, B. Biggio, F. Errica, A. Micheli, F. Scarselli, M. Bianchini, L. Demetrio, P. Bongini, A. Tacchella, et al., Towards learning trustworthily, automatically, and with guarantees on graphs: An overview, Neurocomputing 493 (2022) 217–243.

[8] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, S. Rinzivillo, Benchmarking and survey of explanation methods for black box models, Data Mining and Knowledge Discovery (2023) 1–60.

[9] D. Gunning, D. Aha, Darpa's explainable artificial intelligence (xai) program, AI magazine 40 (2019) 44–58.

[10] Z. Ying, D. Bourgeois, J. You, M. Zitnik, J. Leskovec, Gnnexplainer: Generating explanations for graph neural networks, Advances in neural information processing systems 32 (2019).

[11] L. C. Magister, D. Kazhdan, V. Singh, P. Liò, Gcexplainer: Human-in-the-loop concept-based explanations for graph neural networks, arXiv preprint arXiv:2107.11889 (2021).
[12] H. Yuan, H. Yu, S. Gui, S. Ji, Explainability in graph neural networks: A taxonomic survey, IEEE Transactions on Pattern Analysis & amp; Machine Intelligence 45 (2023) 5782–5799. doi:10.1109/TPAMI.2022.3204236.
[13] J. Kakkad, J. Jannu, K. Sharma, C. Aggarwal, S. Medya, A survey on explainability of graph neural networks, arXiv preprint arXiv:2306.01958 (2023).
[14] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, P. S. Yu, A comprehensive survey on graph neural networks, IEEE Transactions on Neural Networks and Learning Systems 32 (2021) 4–24. doi:10.1109/TNNLS.2020.2978386.
[15] C. Gallicchio, A. Micheli, Graph echo state networks, in: The 2010 International Joint Conference on Neural Networks (IJCNN), 2010, pp. 1–8. doi:10.1109/IJCNN.2010.5596796.
[16] A. Longa, S. Azzolin, G. Santin, G. Cencetti, P. Liò, B. Lepri, A. Passerini, Explaining the explainers in graph neural networks: a comparative study, arXiv preprint arXiv:2210.15304 (2022).
[17] D. Luo, W. Cheng, D. Xu, W. Yu, B. Zong, H. Chen, X. Zhang, Parameterized explainer for graph neural network, Advances in neural information processing systems 33 (2020) 19620–19631.
[18] M. Fontanesi, A. Micheli, M. Podda, Xai and bias of deep graph networks, in: Proceedings of the 32th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning Intelligence (ESANN 2024), 2024. URL: https://doi.org/10.14428/esann/2024.ES2024-85.
[19] M. Fontanesi, A. Micheli, M. Podda, Relating explanations with the inductive biases of deep graph networks, in: Accepted at AixiA 2024 and under publication in the AixiA Springer LNAI, 2024.
[20] M. Fontanesi, A. Micheli, M. Podda, D. Tortorella, Analyzing explanations of dgns through node centrality and connectivity, in: Accepted at Discovery Science 2024 and under publication in the Discovery Science Springer LNCS, 2024.
[21] K. Xu, W. Hu, J. Leskovec, S. Jegelka, How powerful are graph neural networks?, arXiv preprint arXiv:1810.00826 (2018).
[22] C. Morris, M. Ritzert, M. Fey, W. L. Hamilton, J. E. L. et al., Weisfeiler and leman go neural: Higher-order graph neural networks, Proceedings of the AAAI Conference on Artificial Intelligence 33 (2019). doi:10.1609/aaai.v33i01.33014602.
[23] G. Corso, L. Cavalleri, D. Beaini, P. Liò, P. Veličković, Principal neighbourhood aggregation for graph nets, Advances in Neural Information Processing Systems 33 (2020).
[24] P. E. Pope, S. Kolouri, M. Rostami, C. E. Martin, H. Hoffmann, Explainability methods for graph convolutional neural networks, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10764–10773. doi:10.1109/CVPR.2019.01103.
[25] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: International conference on machine learning, PMLR, 2017, pp. 3319–3328.
[26] L. Katz, A new status index derived from sociometric analysis, Psychometrika 18 (1953) 39–43. doi:10.1007/BF02289026.
[27] M. Fiedler, Algebraic connectivity of graphs, Czechoslovak Mathematical Journal 23 (1973) 298–305. doi:10.21136/CMJ.1973.101168.
[28] M. Rathee, T. Funke, A. Anand, M. Khosla, Bagel: A benchmark for assessing graph neural network explanations, arXiv preprint arXiv:2206.13983 (2022) 1–20.
[29] M. Defferrard, X. Bresson, P. Vandergheynst, Convolutional neural networks on graphs with fast localized spectral filtering, Advances in neural information processing systems 29 (2016).
[30] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, et al., Graph attention networks, stat 1050 (2017) 10–48550.
[31] W. Lin, H. Lan, B. Li, Generative causal explanations for graph neural networks, in: International Conference on Machine Learning, PMLR, 2021, pp. 6666–6679.
[32] J. Tan, S. Geng, Z. Fu, Y. Ge, S. Xu, Y. Li, Y. Zhang, Learning and evaluating graph neural network explanations based on counterfactual and factual reasoning, in: Proceedings of the ACM Web Conference 2022, 2022, pp. 1018–1027.