Determining motif explanations for learning on graphs

Bettina Soós[1], Çiçek Güven, Gonzalo Nápoles, and Pieter Spronck

Tilburg University, Tilburg, The Netherlands
B.Soos@tilburguniversity.edu

1 Motivation

Graph learning applications on complex networks often involve high-stakes decisions especially in domains like healthcare, finance or social networks. That is why explainability is considered relevant. A subgraph of a graph is known to explain predictions on the graph [6]. Explainability for a graph learning task often involves highlighting an explanatory sub-graph, consisting of edges contributing the most to a certain outcome. However, finding the explanatory subgraph with arbitrary structure may be less meaningful then finding the recurring pattern in the graph. This work stems from the question if the entire explanatory subgraph, or the motifs representing the isomorphic subgraphs that explains the prediction. Since the motif can break down the explaining subgraph, it provides for a more detailed explanation. At the same time, motif-based explanations compress the information in the larger explanatory subgraph to smaller motifs, providing for sparser explanations.

1.1 Subgraph motifs

A motif $\mathscr{M}$ is a set of all graphs, $G = (\mathscr{V}, \mathscr{E})$, $\mathscr{E} = (u,v)|u,v \in \mathscr{V}$ of specific labeling of the nodes in $\mathscr{V}$ that are isomorphic to the motif. Then, $\mathscr{M} = \{G\}$ and for every $G \in \mathscr{M}$ there is a mapping $m : (u,v) \longleftrightarrow (m(u), m(v))$ for every $(u,v) \in \mathscr{E}$ and $(m(u), m(v)) \in \mathscr{E}_M$, $(\mathscr{V}_M, \mathscr{E}_M) \in \mathscr{M}$ is some graph of the motif.

A motif in a graph $\mathscr{M}_G$ [5][1][3], is understood as all subgraphs that are isomorphic to the motif in the graph. Possible subgraphs determine the set of motifs defined on the graph $\{\mathscr{M}\}_G$. A motif $\mathscr{M}_G \in \{\mathscr{M}\}_G$, projected on a graph covers the graph by the union of subgraph isomorphic to it $G_M = \bigcup_i S_i | \forall S_i \subset G \wedge S_i \in \mathscr{M}_G$, G_M is the graph's covering subgraph for the motif.

1.2 Motif explanations

Given a machine learning model which inputs a graph G_i, and a ground truth value corresponding to the graph y_i that the model predicts, the input graph can be altered and observations can be made on the change in prediction output and performance. Perturbations that are smaller but makes the output more different or be closer to the truth value thought to explain the prediction better. There are several ways to perturb the input by a motif. Examples are to use the motif's covering subgraph G_M, and remove structural components or alter the attributes of the graph that are in the subgraphs, G_M and $G_{\bar{M}} = G - G_M$.

If motif explanations are suitable for the machine learning problem, e.g., if subgraph patterns are relevant for the prediction on the particular problem, motifs can explain 1) if the explanations are optimal, why and when does a machine learning performs well, 2) and given the machine learning model is optimal, what is the most relevant information of the input that is possibly more interpretable than the function implemented by the machine learning model.

While the methodology is independent from the machine learning model, neural networks' prediction are typically more difficult to explain; what is the reason for a model's good performance? Can we tell when the performance is good? What is the most concise representation of the input that contains most of the information that is required for making accurate predictions for the specific problem? Are typically difficult to answer in the context of neural networks.

2 Methodology

The explanatory power of single motifs are determined first, taking two motifs randomly selected as examples (shown in Figure 1). Prediction error is measured by the absolute difference $dy_G = |\hat{y}_G - y|$ of the prediction to the original label. A motif's effect is measured by the difference in the prediction on the perturbed graph to the original prediction $y_P - y_G = \hat{y}_P - \hat{y}_G$ with P perturbation.

The motif is projected on the graph by finding all subgraph isomorphs of the motif in the graph [4], and constructing the union of the subgraphs. Either the union of the subgraph isomorphs, i.e., the covering subgraph $G_M \equiv M$, or its complement in G, $G_{\bar{M}} \equiv \bar{M}$, is used as the perturbed input P for prediction.

Experiments are conducted with three machine learning models, all are custom attention networks [2] tailored to the dataset (a, a2, and b at Table 1). Two are similar networks but (a) was trained for 200 epochs and (a2) for 15. (b) is used with different internal activation functions and number of layers compared to (a) and (a2) and shows different qualitative output for learning and prediction.

The chosen dataset, ogbg-molhiv [7], has a difficult machine learning problem defined on it, meaning that overall it is difficult to achieve very good performance on the dataset, despite it being fairly large in the number of feature dimensions and possibly containing enough information for making accurate predictions. The machine learning task is to predict whether a molecule will inhibit HIV, and there is a large class imbalance with having more negative instances. However, this kind of dataset can benefit from finding subgraph patterns that explain why some positive instances can be predicted well.

3 Results

Motif perturbations have larger effect on the models trained longer (Table 1, Model *a* compared to *a2*). Removing parts of the input graph increases prediction accuracy on the positive class of these models. Motif perturbation has different effect on the model performing differently (Table 1, Model *b*), removing part of the original graph decreases the prediction accuracy of the positive class.

Perturbation both by M and $\bar{M}$ has its effect to the same direction in all cases, in almost all cases this is larger of M. Better motif explanations should have larger (or smaller, for negative class) prediction difference with the original graph $y_M - y_G$ than of their complement graph $\bar{M}$ in G. That is because an explanation contains the most relevant information if the least is in everything that is not part of the explanation. If removing a motif $\bar{M}$ explains the prediction as well as the motif M, then M is not explanatory. Motifs with large (or small) $y_M - y_G$ and small (large) $y_{\bar{M}} - y_G$ can be good explanations.

Given that finding all possible subgraph in a graph is already a computationally challenging problem, evaluating all possible motifs are infeasible. Such constraint, as large $y_M - y_G$ and small $y_{\bar{M}} - y_G$, suggests to be used to find explanatory motifs through combinatorial optimization.

Model	Prediction error $\langle\cdot\rangle_{G_i}$	dy_G	Motif 1 dy_M	Motif 1 $dy_{\bar{M}}$	Motif 1 $y_M - y_G$	Motif 1 $y_{\bar{M}} - y_G$	Motif 2 dy_M	Motif 2 $dy_{\bar{M}}$	Motif 2 $y_M - y_G$	Motif 2 $y_{\bar{M}} - y_G$
a	Class 1	0.5310	0.5294	0.5296	0.0018	0.0015	0.5290	0.5306	0.0021	0.0005
	Class 0	0.4689	0.4705	0.4706	0.0016	0.0018	0.4708	0.4695	0.0019	0.0006
a2	Class 1	0.5504	0.5504	0.5504	1e−5	1e−7	0.5504	0.5504	1e−5	1e−6
	Class 0	0.4496	0.4496	0.4496	1e−5	1e−5	0.4496	0.4496	1e−5	1e−6
b	Class 1	0.9530	0.9994	0.9769	-0.0397	-0.0239	0.9972	0.9613	-0.0330	-0.0083
	Class 0	0.0512	0.0045	0.0138	-0.0473	-0.0374	0.0084	0.0375	-0.0409	-0.0138
	$\lvert\mathscr{V}\rvert, \lvert\mathscr{E}\rvert$	28, 62	21, 33	20, 29			17, 27	22,37		

Table 1: Table shows averaged values of prediction difference dy and values of the difference in prediction $y_P - y_G$ on the perturbed graphs P ($= M$ or $\bar{M}$) and on the original graphs G. Two motifs (Motif 1a and Motif 1b in Figure 1) are used for the perturbations separately, providing the covering subgraphs $M \subset G$ and their complement in G, $\bar{M}$. $\lvert\mathscr{V}\rvert$ and $\lvert\mathscr{E}\rvert$ show average node and edge values of the input graphs.

Summary. In this extended abstract it was demonstrated how to determine the explanatory power of single motifs. Such research is an important forerunner to determine the explanatory power of sets of motifs, and initiates the search for methodology to find particular motifs that are optimized for their explanatory power. It also demonstrates the usefulness of motif explanations to evaluate machine learning model's performance, suggesting how it is possible to explain the machine learning problem itself by the motifs ultimately.

References

1. Yael Artzy-Randrup, Sarel Fleishman, Nir Ben-Tal, and Lewi Stone. Comment on "network motifs: Simple building blocks of complex networks" and "superfamilies of evolved and designed networks". *Science (New York, N.Y.)*, 305:1107; author reply 1107, 09 2004.

(a) Motif 1 (b) Motif 2

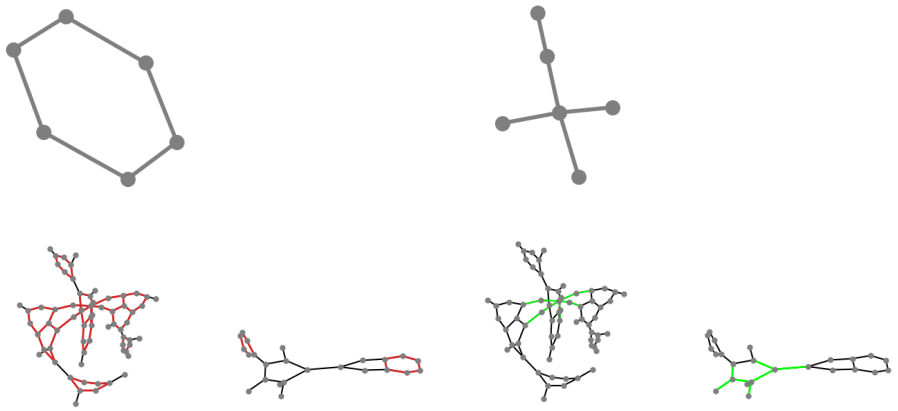

Fig. 1: Motifs and their covering subgraph on two examples from the graph dataset. The dataset (validation set of ogbg-molhiv) contains 4112 graphs, from what Motif 1a was contained in 2977 graphs and Motif 1b was contained in 1661. The motifs are arbitrarily selected and used to conduct the experiments and generate results of Table 1.

2. Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.
3. Piers J Ingram, Michael PH Stumpf, and Jaroslav Stark. Network motifs: structure does not determine function. *BMC genomics*, 7:1–12, 2006.
4. Alpár Jüttner and Péter Madarasi. Vf2++—an improved subgraph isomorphism algorithm. *Discrete Applied Mathematics*, 242:69–81, 2018. Computational Advances in Combinatorial Optimization.
5. Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.
6. Xiang Wang, Yingxin Wu, An Zhang, Fuli Feng, Xiangnan He, and Tat-Seng Chua. Reinforced causal explainer for graph neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2297–2309, 2022.
7. Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.