Tracking number : 839-1841

Original Article Title: Towards Semantic Understanding of GNN Layers Embedding with Functional Semantic Activation Mapping

Dear Editor,

We thank all reviewers for their valuable feedback and comments. We sincerely appreciate the opportunity to revise and resubmit our manuscript titled "Towards Semantic Understanding of GNN Layers Embedding with Functional Semantic Activation Mapping". We are grateful to the reviewers for their insightful comments, which have significantly strengthened our work. Please find enclosed:

1. A point-by-point response to all reviewer comments (attached below)
2. Revised manuscript with changes highlighted in green ("NeSy.FSAM_highlighted.pdf")
3. Clean version of the revised manuscript ("NeSy.FSAM_Cleaned.pdf") without highlights.

Thanks and best regards
Kislay et al.

Reviewer 1, Concern # 1:

- Page 2 Lines 2-3 missing . at the end of sentence

- Page 2 Line 9: The oversmoothing has been studied -> could be rephrased as "Oversmoothing has been studied" or "The oversmoothing phenomenon"

- Page 3 Line 41 punctuation issue. "are then combined into logic formulas. as it learns concepts" -> "are then combined into logic formulas, as it learns concepts"

- Ensure that Section and Figures are always referenced consistently (e.g. 5.3 instead of Section 5.3 at page 6 lines 32-38)

- Table 2. It is not clear how the the caption was revised to explicitly state that the table compares the final accuracy of models with different depths rather than the accuracy computed at different layers of a single model.

Author response: Thank you for your valuable comment. We have thoroughly checked and corrected all typographical and grammatical errors in the revised manuscript.

Reviewer 2, Concern # 1: Please explain clearly the setting used in Section 3. For example, are all definitions with respect to a fixed graph, or to a class of graphs? What kinds of graphs are allowed, and with what features (Boolean, real?) What is the task being solved (node classification, graph classification?) and what kinds of inputs and outputs do you expect?

Author response: We greatly appreciated your feedback. In the revised manuscript, Section 3.1 has been expanded to clarify the graph assumptions, feature domains, task type, and expected inputs and outputs. All the changes are highlighted in green.

Reviewer 2, Concern # 2: In the Introduction, please provide a summary of FSAM and the kind of output that it produces. The difficulty in understand what, precisely, is the output of FSAM (and why do we call it "semantic") significantly hinders the rest of the paper. A simple example would be extremely helpful.

Author response: We thank the reviewer for this valuable suggestion. In the revised manuscript (page 2, lines 3–12), we have added a concise summary of FSAM, including a clear description of its outputs and the rationale for the term "semantic", accompanied by a simple example. Furthermore, the graphs in Figures 13-16 show the semantic relationships inferred from the activation patterns at different layers, which are referred to as functional-semantic graphs.

Reviewer 2, Concern # 3: Please give a more complete description of the experiment set-up and the metrics that are used. Many important definitions are missing, which makes it difficult to corroborate or object to the conclusions drawn in Section 5 (see detailed comments below). If space is an issue, I would recommend shortening the current discussion of the experiments. For example, the descriptions in lines 1-12 of p.9 or 21-27 of p.9, 31-49 of p.9 are somewhat repetitive, and similarly for other subsections.

Author response: In Section 5 (subsections 5.1 and 5.2), we have summarised the key experimental details and findings while preserving all important information. The experimental setup, metrics, and definitions have been expanded for completeness and redundant descriptions have been removed to improve clarity and readability in the revised manuscript.

I have some additional questions, which I believe need not be addressed in a revision of the paper, but could potentially reinforce its main message. For example, the authors study how interpretability diminishes as the number of layers increases. Would the same phenomenon occur if we increase the dimension of the hidden layers, instead of the number of layers? Another interesting experiment would be to apply FSAM at different points during training. I'd like to see loss curves and study potential relations between the phenomenon described in this paper and grokking.

We appreciate these thoughtful suggestions. While they are beyond the current scope of this study, we recognise their potential to deepen our understanding of the phenomena observed. Investigating whether increasing hidden layer dimensionality produces similar interpretability degradation, as well as applying FSAM at different stages of training to capture temporal evolution, are indeed promising directions. We also agree that tracking loss curves and examining potential links to phenomena such as grokking could provide valuable complementary insights. These aspects have now been added as part of our planned future work.

Reviewer 2, Concern # 4: P1, l21: "identifying locally relevant subgraphs (instance-level explanations)" This seems to be conflating two different aspects of explanation. Focusing on locally relevant subgraphs (unless I misunderstood what this means) is not necessarily an instance-level explanation. For example, a model could predict "True" for all vertices that are part of a 3-clique (on any given input graph). This refers to a locally relevant subgraph, but it is not instance-based, since it is a general rule that applies to any input.

Author response: We acknowledged the potential ambiguity in conflating locally relevant subgraphs with instance level explanations. In the revised manuscript (Abstract, p.1, lines 18–23), we have clarified that our reference is specifically to instance level explanations defined as explanations that aim to justify a model's prediction for a given input.

Reviewer 2, Concern # 5: P2, l3: "behaviour In" -> "behaviour. In"

State of the Art: you might wish to consider a brief discussion of works that provide explanations of GNN predictions [e.g. [1]] by extracting logical rules implied by the network, or by translating the entire GNN into a set of rules. You could argue that the FSAM approach addresses an important gap in rule-based GNN explanation literature (namely, the extracted rules are too complex), which further highlights the significance and timeliness of FSAM analysis.

[1] Tena-Cucala, D. and Cuenca Grau, B. Bridging Max Graph Neural Networks and Datalog with Negation. Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning.

Author response: In the revised manuscript, we have incorporated a discussion of recent rule-based approaches for explaining GNN predictions, including the work of Tena-Cucala and Cuenca Grau (2024), in the State of the Art section (p.3, lines 29–39).

Reviewer 2, Concern # 6: P5, l26: "n represents the number of neurons […] corresponding either to the number of nodes in G" – does this mean that the size of the GNN depends on the size of a specific input? This seems strange. One of the advantages of GNNs is precisely that they can be defined independently of any input and can be applied to graphs of any size.

Author response: We acknowledged the ambiguity in the original statement and have revised the Mathematical Formulation section (section 3.1 has been extended) to clarify that the number of neurons in each layer (h_i) is an architecture defined parameter, independent of the size of the input graph (n). The revised formulation now clearly separates graph dependent quantities from model parameters avoiding the earlier misinterpretation.

Reviewer 2, Concern # 7: P5, l49: what is the rank of an activation value? If I am following along, a_i and a_j are scalars. Also, what does standard deviation here refer to? Are you assuming a distribution over a set of input graphs? If so, I would suggest making this explicit.

Author response: In the revised manuscript, we have made it clear that a_i refers to the vector of activation values for neuron i across all n nodes in the fixed transductive setting. The operation rank(.) generates the rank transformed activation vector, with ties handled according to the standard Spearman correlation procedure. The standard deviation σrank(ai) is calculated over the ranked activation values for the n nodes in the given graph rather than across multiple input graphs. This clarification has been added in Section 3.1 (page 6, lines 3-14).

Reviewer 2, Concern # 8: P5, l50: why do activations have monotonic dependencies? Couldn't they be non-monotonic due to negative parameters in W^(\ell)?

Author response: We agree that neuron activations can indeed be non-monotonic functions of each other due to the presence of signed weights and nonlinear activation functions. In the revised manuscript, we have clarified that Spearman's p is used because it captures any consistent rank order relationship whether increasing or decreasing across nodes, without requiring the functional relationship between neurons to be strictly monotonic. This property makes it particularly suitable for identifying correlation patterns in activations, even when the underlying mapping between neurons is non-monotonic. This clarification has been added into Section 3.1 (Page 6, Lines 30–38).

Reviewer 2, Concern # 9: P6, l1: what are activations *from different classes*? Does this refer to activations of the same neuron for a collection of inputs that receive the same classification, in a classification setting?

Author response: In the revised manuscript, Section 3.1 (Page 6, Lines 40–47) clarifies that activations from different classes refer to the activation values of the same neuron evaluated across all nodes belonging to different ground-truth classes in the dataset (not predicted classes).

Reviewer 2, Concern # 10: P6, l3: "activations co-escalate beyond a certain depth" – what does this mean? Does this mean that they increase sharply with respect to the number of layers in the network (across different GNNs)? Or that for the same model, co-activations of deeper layers tend to be higher than co-activations of shallow layers, or co-activations of a deep layer and a shadow layer?

Author response: In the revised manuscript, Section 3.1 (Page 6, Lines 49–51, Page 7, lines 1-4), we clarify that co-escalation refers to the depth dependent increase in pairwise spearman correlations between neurons within the same model, with deeper layers showing higher correlations especially between neurons associated with different ground-truth classes than shallower layers. It indicates a loss of class specific representation at greater depths.

Reviewer 2, Concern # 11: P6. L5: what are the binary input variables? Are we assuming that each input vertex has a Boolean value 0 or 1? n_1 and n_0 seem to be the number of inputs "0" and the number of inputs "1", but it would be worth saying this explicitly.

Author response: In the revised manuscript, Section 3.1 (Page 6, Lines 44–48), we explicitly clarify that the binary input variable in the point biserial correlation computation is the class indicator vector where nodes belonging to the target class are assigned a value of 1, and all others are assigned 0. This grouping is performed in a one-versus-rest manner for each class and is independent of whether the original node features are binary or continuous. Here, n1 and n0 denote the number of nodes assigned values 1 and 0, respectively.

Reviewer 2, Concern # 12: P.6. l13-19: I could not follow this explanation. Maybe it's best to have this later in the paper, or to provide a single example. At this point, it's not clear what "affect the semantics structure" or "provide minimal information" means.

Author response: In the revised manuscript, Section 3.1 (Page 7, Lines 1–10), we have clarified the explanation and provided a concrete interpretation. Specifically, "affect the semantic structure" refers to changes in the arrangement and separability of class specific activation patterns in the learned representation space, as visualised through FSAM graphs. "Provide minimal information" means that deeper layers contribute little new class discriminative signal, as evidenced by increased overlap in activation distributions across classes. We also give an example illustrating how beyond a specific depth neurons representing different classes exhibit similar activation patterns, reducing the model's ability to distinguish them.

Reviewer 2, Concern # 13: P6. L32: As pointed out before, it is hard to understand this contribution without having a good idea of they kind of "semantic structure" shown by FSAM. Similarly, in line 43: "more coherent FSAM representations" it is still not clear what this means.

Author response: In the revised manuscript (Page 7, Lines 23–28, lines 32-35), we clarify that semantic structure refers to the organisation of neurons into communities whose activation patterns align with ground truth class boundaries and relationships. At the same time, more coherent FSAM representations denote FSAM graphs where these communities are compact, well separated and semantically consistent with class labels.

Reviewer 2, Concern # 14: P7, l30: "Semantic graphs" seems a crucial notion to understand FSAM, so its first mention should not be here.

Table 2: this table uses Pearson Correlation, but page 5 states that Spearman correlation is used. Furthermore, how is this correlation exactly computed?

Author response: In the revised manuscript, as noted in our reply to Concern #3, Section 5 has been fully revised, and the concept of semantic graphs is now introduced earlier in Section 3.1 (Page 5, Lines 12–18) to clarify their role in FSAM before their use in later sections.Regarding Table 2 (Page 9), we have added a footnote clarifying that Pearson correlation is used here to measure the linear association between FSAM derived metrics (averaged per model configuration) and model accuracy across different layer depths. This differs from the Spearman correlation in Section 3.1, which is applied to neuron-level activation patterns where monotonic but potentially non-linear relationships are expected. Pearson correlation in Table 2 is computed between the set m_L of averaged FSAM metrics and the corresponding accuracies a_L for L=1,…,4 layers on each dataset.

Reviewer 2, Concern # 15: P10, l14: the definition of communities is rather unclear. Do they include neurons within a layer, or across multiple layers? What allows one to identify communities across layers? Why are communities relevant in this particular experiment?

Author response: We thank the reviewer for the comment. We have clarified the definition of communities in the revised manuscript (page 10, lines 35–40). Communities are detected within each GNN layer using the Louvain method on the FSAM coactivation graph. Detection is layer specific, but we track similar activation profiles across layers to study their evolution. Communities are relevant as they represent semantically related neuron groups and their changes with depth reveal effects on representation, accuracy, and oversmoothing.

Reviewer 2, Concern # 16: P10, l25: this again mentions "clusters of semantically related fields" but it is unclear what this means. What are 'semantic relations' in this context?

Why is it important to consider co-activations? Can you provide more intuition for this? Why do we care, specifically, about relations between neurons?

Author response: In the revised manuscript (p.10, l.35–40), addressing Concern #15, we clarify that clusters of semantically related fields refer to groups of neurons whose activation patterns across nodes are highly correlated, indicating they respond similarly to specific structural or feature patterns in the graph.