Improvement of Attention Mechanism Explainability in Prediction of Chemical Molecules' Properties

Bartosz Durys, Arkadiusz Tomczyk[0000−0001−9840−6209]

Lodz University of Technology
Institute of Information Technology
al. Politechniki 8, 93-590 Łódź, Poland
bartdurys@gmail.com, arkadiusz.tomczyk@p.lodz.pl

DOI:10.34658/9788366741928.17

Abstract. *In this paper, the analysis of selected graph neural network operators is presented. The classic Graph Convolutional Network (GCN) was compared with methods containing trainable attention coefficients: Graph Attention Network (GAT) and Graph Transformer (GT). Moreover, which is an original contribution of this work, training of GT was modified with an additional loss function component enabling easier explainability of the produced model. The experiments were conducted using datasets with chemical molecules where both classification and regression tasks are considered. The results show that additional constraint not only does not make the results worse but, in some cases, it improves predictions.*

Keywords: *attention mechanism, graph transformer, graph neural network, explainability, chemical molecules*

1. Introduction

Graphs for a long time were an uncommon data type in machine learning solutions. Lately, many new methods for graph prediction tasks have been proposed, including those utilizing attention-based graph neural network operators. Interpretability of the attention mechanism in natural language processing problems is a well-researched subject. In this work, we investigate it in the context of the graph data structures. We evaluate the performance of the chosen operators on selected benchmark graph datasets containing chemical molecules. An original contribution of this work is a mechanism that forces attention coefficients to be more precise in indicating, which neighbouring nodes, and consequently which relations, are particularly important for prediction. The obtained outcomes reveal that, although the used mechanism imposes additional constraints on the trained neural network, it surprisingly does not aggravate the prediction results increasing, at the same time, model explainability.

2. Method

Three Graph Neural Network (GNN) operators were selected: a classic Graph Convolutional Network [1] (GCN), a more sophisticated Graph Attention Network [2] (GAT) and a Transformer's adaptation called Graph Transformer [3] (GT). The last two of them use the attention mechanism. The working principle (transformation of node embeddings **h** in layer t) for above operators can be summarized as:

$$\mathbf{h}_i^{t+1} = \sigma\left(\alpha_{ii}(\mathbf{W}_1\mathbf{h}_i^t + \mathbf{b}_1) + \sum_{j\in\mathcal{N}(i)} \alpha_{ij}(\mathbf{W}_2\mathbf{h}_j^t + \mathbf{b}_2)\right) \tag{1}$$

which, although is not the most general formulation for all GNN operators, is sufficient for further considerations. In this formula σ represents non-linear activation function, matrices $\mathbf{W}_1$, $\mathbf{W}_2$ as well as vectors $\mathbf{b}_1$, $\mathbf{b}_2$ are (if present) directly trainable parameters and $\mathcal{N}(i)$ denotes set of nodes connected with given node i. Coefficients α are fixed in GCN and depend on graph structure only. In GAT and GT these are indirectly trainable attention coefficients that take into account embeddings of connected nodes. In both cases for a given node i those coefficients are normalized with softmax function, which means that $\alpha_{ij} \in [0, 1]$ for $j \in \mathcal{N}(i)$ and their sum is equal to 1.

Training GT model it can be frequently observed that for a given node i coefficients α_{ij} tend to have similar values. It means that all the neighbouring nodes have similar influence on the calculated embedding of the node i and, consequently, it does not allow to draw any conclusions explaining the final predictions. To make those attention coefficients more interpretable we have introduced a new loss function component:

$$\mathcal{L}^{explain} = \sum_i \left(1 - \max_{j\in\mathcal{N}(i)} \alpha_{ij}\right) \tag{2}$$

It forces the model to direct its attention to only one neighbour while aggregating embeddings from each and every node. This component utilizes the softmax normalization of α_{ij} for a given i since optimally there should be only one 1 value among α_{ij} and the rest of them should be equal to 0. The final loss function used during training was the following:

$$\mathcal{L} = \mathcal{L}^{prediction} + \lambda \cdot \mathcal{L}^{explain} \tag{3}$$

where the first component was dependent on the considered task and it was MSE for regression and cross-entropy for classification. The parameter λ controls trade-off between those components, and it has been experimentally set to 0.1.

For experiments we have chosen two widely-used sources of chemical graph data – MoleculeNet [4] for graph-level regression tasks and TUDataset [5] for graph-level classification tasks. From each data source, we have selected three

Table 1: Quality metrics for graph-level regression tasks from MoleculeNet.

Dataset	Operator	Validation set			Test set		
		MSE	**Standard deviation**	**Best**	**MSE**	**Standard deviation**	**Best**
ESOL	GCN	35.24	9.11	8.15	35.24	8.80	8.11
	GAT	**11.76**	**5.09**	4.62	**11.01**	**5.04**	**4.46**
	GT	32.07	18.40	5.61	31.58	17.80	5.41
	GT with $\mathcal{L}^{explain}$	17.87	11.29	**4.53**	17.25	11.11	**4.46**
FreeSolv	GCN	78.64	38.21	17.42	75.11	34.37	17.01
	GAT	36.02	16.94	14.42	**33.95**	**15.02**	14.17
	GT	44.30	26.35	**13.91**	41.87	25.38	14.01
	GT with $\mathcal{L}^{explain}$	**34.71**	**11.35**	14.14	34.11	17.25	**13.86**
Lipophilicity	GCN	12.65	6.75	2.34	12.69	6.84	2.35
	GAT	10.67	4.54	2.12	10.64	4.58	2.16
	GT	12.51	9.19	2.44	12.44	9.15	2.46
	GT with $\mathcal{L}^{explain}$	**2.62**	**0.80**	**1.62**	**2.56**	**0.77**	**1.64**

datasets: ESOL (prediction of water solubility), FreeSolv (estimation of hydration free energy) and Lipophilicity (finding octanol/water distribution coefficient) from MoleculeNet and AIDS (identification of molecule's activity against HIV), ENZYMES (assign a molecule to one of the six Enzyme Commission top-level classes) and PROTEINS (prediction if a protein is an enzyme) from TUDataset. In MoleculeNet feature vectors for each dataset contained nine numerical features describing atoms, e.g. its hybridization, while in TUDataset it was a one-hot encoded representation of node class (chemical element in AIDS or secondary structure element in ENZYMES and PROTEINS).

To train and evaluate the performance of GNN operators on datasets, we have split each dataset into training, validation, and test sets using an 80/10/10 proportion. To ensure a fair comparison and easier interpretability, we have limited our research to only single-headed attention mechanisms. For each operator, we have selected two GNN layers with batch normalization, dropout and ReLU as activation function σ. After that, the global average pooling was used to aggregate the calculated hidden node embeddings. Finally, we have used an MLP to generate our final predictions. We have repeated every experiment 50 times with 500 epochs per repeat, and averaged the results using the best epoch on the validation set. This approach allowed us to obtain reliable and robust results for each dataset and operator combination.

3. Results

Starting the analysis of the results from MoleculeNet's datasets in the Table 1, we can observe several interesting phenomena. First, which is expected, we can see that operators with the attention mechanism perform better than simple GCN. However, the GT operator suffers from a high standard deviation value, which in-

Table 2: Quality metrics for graph-level classification tasks from TUDataset.

Dataset	**Operator**	**Validation set**			**Test set**		
		Accuracy	**Standard deviation**	**Best**	**Accuracy**	**Standard deviation**	**Best**
AIDS	GCN	80.22	2.81	**88.50**	79.84	**2.27**	84.00
	GAT	79.84	**2.74**	85.50	79.91	2.55	86.00
	GT	**81.05**	2.80	86.00	**80.49**	2.46	**87.00**
	GT with $\mathcal{L}^{explain}$	79.75	2.76	86.00	80.03	2.42	86.00
ENZYMES	GCN	30.90	4.14	43.33	20.87	5.69	35.00
	GAT	32.03	4.79	**46.67**	21.43	**5.56**	35.00
	GT	35.17	**3.45**	43.33	24.03	5.80	**40.00**
	GT with $\mathcal{L}^{explain}$	**36.77**	3.81	45.00	**24.33**	5.74	35.00
PROTEINS	GCN	73.69	3.99	81.08	68.88	5.27	78.57
	GAT	73.39	**3.43**	81.08	68.55	4.63	76.79
	GT	**74.29**	3.97	**81.98**	**69.20**	5.04	**79.46**
	GT with $\mathcal{L}^{explain}$	74.22	3.64	**81.98**	68.66	**4.50**	78.57

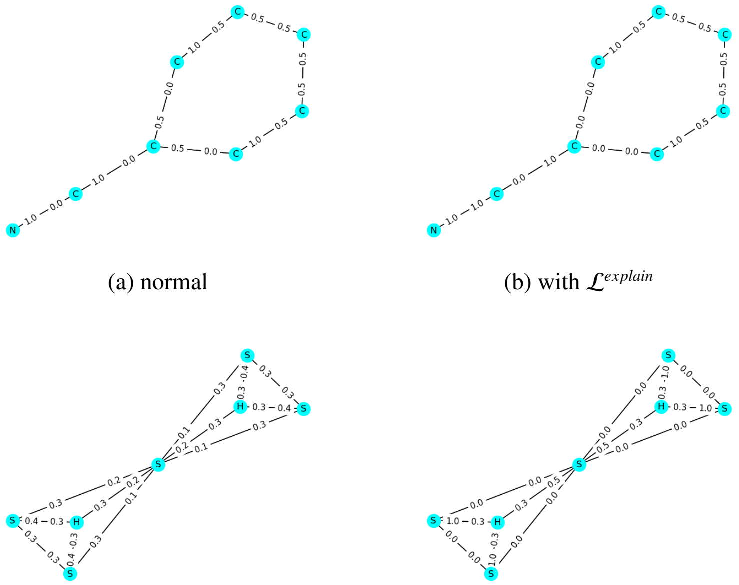
dicates that it is difficult to train, much like the original Transformer. Surprisingly, adding the $\mathcal{L}^{explain}$ function significantly improves the training of the model.

When it comes to the evaluation of TUDataset, our results are close in value to each other. As shown in the Table 2, the GT operator performs slightly better overall. The biggest difference can be seen in the ENZYMES dataset, which has six classes, whereas the other problems are binary classifications. These results suggest that the GT operator may be a better choice for classifications with a large number of classes.

To show the impact of $\mathcal{L}^{explain}$ component, we have prepared visualisations of attention coefficients, which are shown in Figure 1. Values near each node represent weights during the aggregation of its neighbours. In the first figure, we can observe that the proposed modification made the nitrogen atom more important. In the second figure, the model focused more on helices rather than sheets elements. This behaviour exhibited by the model could be a valuable source of information for explainability. In organic compounds, there are many chemical substituents that can have an impact on the molecules' properties. By utilizing a modified loss function, we may be able to better represent and understand their effects, ultimately leading to improved results, as demonstrated in this paper.

4. Conclusions

In this work, a modification of the training loss function for attention-based models was proposed. Its goal is to improve the interpretability of attention coefficients. Outcomes reveal that indeed it works correctly and, what is more, it does not worsen prediction results. An explanation of this phenomenon can be the fact that network architectures are frequently overdesigned and the proposed constraint allows to select the model with desired properties out of many equivalent (similarly predicting) solutions. The quality of the discussed method was assessed



(a) normal

(b) with $\mathcal{L}^{explain}$

(c) normal

(d) with $\mathcal{L}^{explain}$

Figure 1: Attention visualisation: (a), (b) – benzonitrile from the FreeSolv set with GT normalization coefficients from the first layer. C – carbon, N – nitrogen, (c), (d) – enzyme from the ENZYMES set with GT normalization coefficients from the first layer. H – helix, S – sheet. Source: own work.

using datasets with chemical molecules. It should be, however, emphasized that it can be of use in any task where an attention mechanism is used, leading to better explainability of model behaviour.

References

[1] Kipf T.N., Welling M., *Semi-supervised classification with graph convolutional networks*, *CoRR*, 2017, doi: 10.48550/arXiv.1609.02907.

[2] Veličković P., Cucurull G., Casanova A., Romero A., Liò P., Bengio Y., *Graph attention networks*, 2017, doi: 10.48550/arxiv.1710.10903.

[3] Shi Y., Huang Z., Wang W., Zhong H., Feng S., Sun Y., *Masked label prediction: Unified massage passing model for semi-supervised classification*, *CoRR*, 2020, doi: 10.48550/arxiv.2009.03509.

[4] Wu Z., Ramsundar B., Feinberg E., Gomes J., Geniesse C., Pappu A.S., Leswing K., Pande V., *Moleculenet: a benchmark for molecular machine learning*, *Chem. Sci.*, 2018, vol. 9, pp. 513–530, doi: 10.1039/C7SC02664A.

[5] Morris C., Kriege N.M., Bause F., Kersting K., Mutzel P., Neumann M., *Tudataset: A collection of benchmark datasets for learning with graphs*, *CoRR*, 2020, doi: 10.48550/arXiv.2007.08663.