Research Statement

The benefit of graph data is the ability to model complex data relationships that are deeply structured to suit inferences about indirect facts and indirectly related information. Edges are just as influential and complex as vertices/nodes, making graphical representation learning to emerge as a practical approach to graph data exploration in recent years. It has been significant in the transition of graph vertices, edges, or subgraphs into low-dimensional dense embeddings that maintain the structural features and attributes of the graphs. Retained embeddings can then serve as input to standard machine learning models for downstream tasks. However, Graph Neural Networks (GNN) research has faced obstacles due to the very limited data sizes of traditional graph datasets after several years of development. Therefore, *G*raph Data Augmentation (GDA) enjoys widespread acceptance and has rapidly improved performance for diverse GNN. Their success is based on two integral properties of relational approaches:
1. topology-level and
2. feature-level augmentation.

On the other hand, there are many limitations of the proposed methods and their implementations, in particular
~ *Oversmoothing:* its inability to deal elegantly with the small information-to-noise ratio in high-order neighbors, and the high information-to-noise ratio in low-order neighbors; and
~ *Domain adaptation:* the hypothesis is that the most beneficial of augmentation types can be data-specific.

My research statement focuses on extensions of GNN approaches that alleviate the above shortcomings in both node and graph classification tasks.

Research Contribution(s)
SCMvL: A Search-based Contrastive Multi-View Learning for Graph Classification:
Graph contrastive learning (GCL) is obstructive and often yields unsatisfactory performance without data augmentation, as the contrastive loss efficiently estimates the mutual information (MI) between the positive and the negative views.
While past research in this area assumes a particular augmentation works for all datasets, we systematically compose a simple Search-based Contrastive Multi-view Learning (SCMvL) engaging pre-training framework. The approach collectively searches for the best augmentations for multi-view construction to learn intrinsic and transferable structural representations of various datasets by employing contrastive loss for MI estimation. This minimizes the gap between the READOUT anchor and positive information while maximizing that of the positive and negative sampled data.
The main contributions in this area are:
- √ The method takes the source graph Gs and the pre-trained embeddings (Gt and Gt′) as input for downstream task evaluation. The pre-trained embeddings and the source graph are finally fed to GNN variants for graph classification.
- √ We prolong the training mechanism during pre-training for more negative samples for each sample, considering their effectiveness in contrastive learning. It relatively balances the small information-to-noise ratio in high-order neighbors as it generates more negative samples for a more varied split of total datasets.

Extensive experiments on benchmark graph datasets of different sizes with distinct characteristics show that our approach outperforms the compared baselines in 7 out of 10, making better generalizable features and avoiding over-smoothing.

Graph Contrastive Multiview Learning: A Pre-Training Framework for Graph Neural Network

Graph contrastive-based pre-train learning task has benefited from recent developments such as node and graph classification and similarity search. For performance reasons, similarity search, which is usually performed in non-transferable to out-of-domain data, mostly yield low results competitively in various task. In contrast, the hypothesis is that keeping the task within a domain-specific equalizes the alleviation of the extra similarity search in the non-transferable to the out-of-domain functions and can benefit downstream tasks. In this work, we engage a fraction of the datasets for pre-training as they have similar local network topologies in the vector space, making it compatible with the vertices and graph when fine-tuning. The approach saved us the extra effort of the similarity search.

The main contributions in this area are:

- √ We generate pre-train embeddings based on domain-specific tasks for multi-view representation learning to alleviate the extra effort for the similarity search in out-of-domain tasks.
- √ The standard GCN for classifying nodes and graphs reveals that they are usually superficial and limit the number of layers to 2. However, past research believes that the network layer can be extended on a large dataset, where they extended the encoder model to a 3-layer GCN model. Due to the structural similarity and transferability, we can increase the number of layers in our proposed model without over-smoothing in a transductive setting.

Extensive experiments on ten graph classification tasks demonstrate that our method outperforms the compared baselines with current state-of-the-art performance in nine out of the ten datasets.

Future Research Plans

My current research plans are closely tied to the above results, in particular, to resolve several open questions in the area of GNN including, but not limited to the following:

1. While a 3-layer most satisfactory benefits our proposed method, increasing the number of layers beyond that degrades the model's performance and also overfits it in various GNN models. Nevertheless, we noted that performance on some datasets increases with the GCN and GraphSAGE architecture beyond 3 layers. Intuitively, this can be attributed to the immense value of edge density in some datasets. When datasets are more densely connected, their representations become indistinguishable when the model goes deep. Although there is an observable performance gain beyond the 3-layer encoder, the training procedure shows inconsistency, with numerous peaks in both validation loss and accuracy. Furthermore, the GCN model is known for its performance in transductive settings. In contrast, GraphSAGE is known for its generalized aggregator, mini-batch training, and fixed-size neighbor sampling algorithm to accelerate the training process, hence being more scalable in the inductive setting. *Even though our model is not geared toward deeper GNN, can the systematic observation from the work combat overfitting and go deeper with pre-trained embeddings*?
2. As a result of prolonged training for more negative samples for each sample, our approach relatively balances the small information-to-noise ratio in high-order neighbors. However, not a high information-to-noise ratio in low-order neighbors. Therefore, increasing the network layer can cause an interaction between high-order neighbors to bring over-smoothing. *What are the tradeoffs between a small information-to-noise ratio in high-order neighbors and a high information-to-noise ratio in low-order neighbors in fixing the interaction that brings over-smoothing*?
3. *How to devise GNN method for other classification tasks, such as node classification, and beyond self-supervised learning, such as unsupervised, semi-supervised, and reinforcement learning*?

Although the above research statement is focused on graph neural network. I have an exemplary passion and motivation to pursue multidisciplinary research at the intersection of computer vision and natural language processing.