GraphHSCN: Heterogenized Spectral Cluster Network for Long-Range Representation Learning

Camille Dunning, Zhishang Luo, Sirui Tao

Halıcıoğlu Data Science Institute, University of California, San Diego

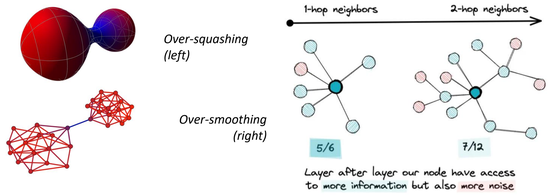
Abstract

Context

Graph Neural Networks (GNNs) have gained tremendous popularity in recent years due to their impressive performance in solving graph-structured data, which is very common in real-world applications such as social network analysis, recommendation systems, drug discovery, traffic prediction, computer vision, natural language processing, and finance.

Challenges

Most graph models only handle local interactions, limited to nodes within a few neighborhoods. Expanding to distant nodes may lead to over-smoothing and over-squashing. To address this, transformer and diffusion models are introduced, but not tested on large graph datasets with long-range interactions.



Goal: We propose GraphHSCN - Heterogenized Spectral Cluster Network, a new MP-based approach to long-range graph modeling.

Dataset

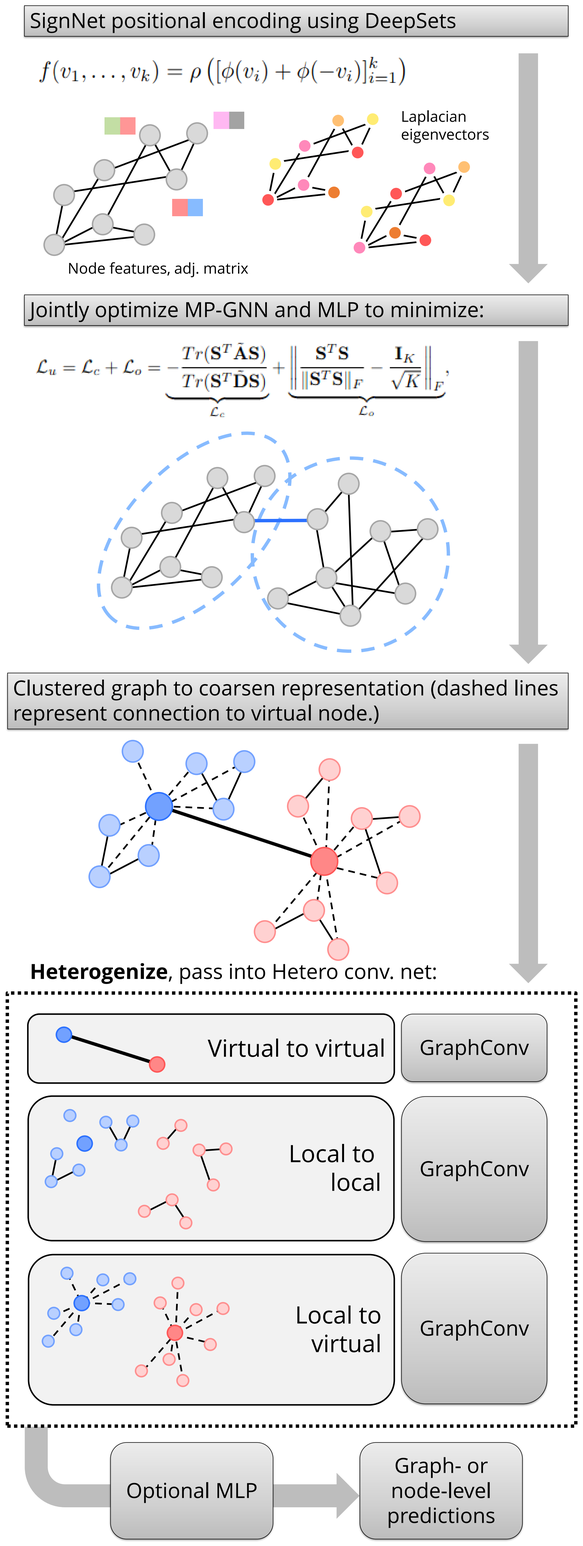
Resampled Citation Networks (single-graph): Our first benchmark considers the transductive semi-supervised node classification task on citation networks. The Cora, CiteSeer, and PubMed networks' node features are bag-of-words representations of documents and edges represent citation links. To build the labeled training dataset, 20 instances of each class are randomly sampled. 1,000 instances are sampled for the test dataset, and the remaining 500 for the validation set. Our benchmarking method repeats training on three different seeds of splitting on the datasets. With the aim of tailoring these datasets to the task of long-range model benchmarking, we adopt the resampling scheme introduced in the Hierarchical Graph Network paper. This scheme retains the process of selecting 20 examples from each class for training, but rather than doing so uniformly at random, for a drawn node, it "sanitizes" its *k*-hop neighborhood of labels. In this study, we employ a buffer of $k = 1$. For future work, we should try for $k = 2$.

Peptides (multi-graph): Peptides-func and Peptides-struct are derived from 15,535 peptides, short chains of amino acids, retrieved from SATPdb. The molecular graph of a peptide is much larger than that of a small drug-like molecule as each amino acid is composed of many heavy atoms. The graphs are constructed such that the nodes correspond to the heavy (non-hydrogen) atoms of the peptides while the edges represent the bonds between them. Both datasets use the same set of graphs but differ in their prediction tasks.

Dataset	# Graphs	# Nodes	Avg. # Nodes / Graph	# Classes	# Node Features	Task
Citeseer	1	3,312	-	6	3,703	NC
Cora	1	2,708	-	7	1,433	NC
PubMed	1	19,717	-	3	500	NC
Peptides-func	15,535	-	150	10	9	GC
Peptides-struct	15,535	-	150	11	9	GR

Methodology

Architecture Diagram



Architecture

In order to improve message-passing neural networks (as opposed to using attention/improving transformer architecture), we designed our architecture to be comprised of three main modules:

(1) SignNet position encoding - not leveraged in our experiments, but we propose this as part of our final architecture, so that each position can be have a unique representation

(2) Spectral clustering model to jointly optimize the minCUT and orthogonal losses

(3) Heterogeneous convolutional network to train on relationships realized from clustering.

SignNet

- O(d) invariant
- Permutation equivariant
- Computational complexity: $O(NKF)$

N is number of nodes, K is K eigenvectors, and F is the hidden dimension of the learned node feature

Spectral Clustering

- minCUT problem formulation:
- relaxed formulation:
- Optimzied by the joint loss $L_u = L_c + L_o$, which approximates a continuous relaxed formulation of the minCUT problem
  - Cut loss $(L_c) = \frac{Tr(S^T\tilde{A}S)}{Tr(S^T\tilde{D}S)}$
  - Orthogonal loss $(L_o) = ||\frac{(S^TS)}{||S^TS||_F} - \frac{I_K}{\sqrt{K}}||_F$

($||\cdot||$ is Frobenius norm and $\tilde{D}$ is the degree matrix of $\tilde{A}$)

- Computational complexity:

It is dominated by the numerator in Lc , which is O(NK(K + N)). As $\tilde{A}$ is usually sparse, we reduce it to O(K(E + NK)) where E is the number of non-zero edges in $\tilde{A}$.

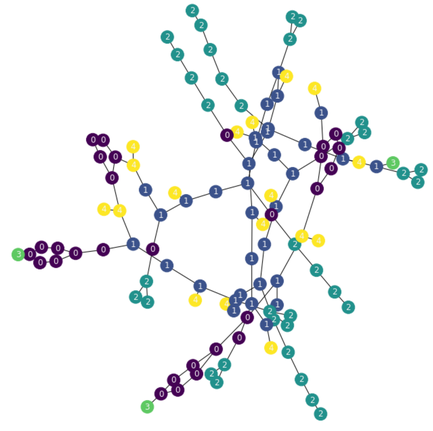


Figure: Spectral Clustering Example for Peptides

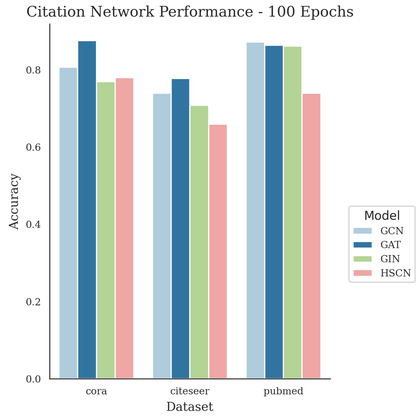
Heterogeneous GNN

- 3 types of connection – (V2V, L2L, L2V)
- Computational complexity: $O(LNF^2)$

L is number of layer and F is the node feature dimension

Results

For the peptides datasets, we report our experiment on five hidden layers, although we also obtained similar results for two layers. Two-layer results are reported for the citation networks. Our architecture well outperforms traditional message-passing networks on graph-level tasks, also converging in less epochs. However, it is still outperformed by the SAN transformer. On node-level tasks, we observe a lesser performance, as the graphs are larger. Particularly, the Citeseer dataset has more features than nodes, so graph coarsening by our architecture could have increased the level of overfitting.



Conclusion

Contribution

Our model achieves state-of-the-art results compared to local attention-based models for predicative ability on common as well as datasets benchmarked for long-range tasks.

Complexity

Our model significantly reduced the computational complexity of SOTA model SAN's O(n^2) to O(nk^2) while reaching a similar level of performance.

Future Work

1. **Attention mechanisms.**
In order to deal with the increasing number of clusters for even larger graph, we will use attention mechanism to reduce the cost of computing the cluster-level connections. This might mitigate the underperformance of Graph-HSCN on the citation network datasets.
2. **Test on more datasets.**
We planned to test the remaining datasets from "Long Range Graph Benchmark" and the circuit board design datasets from Qualcomm.
3. **Run SignNet PE after Spectral Clustering.**
More ablation studies are also needed to see if moving certain components around will affect the performance or space & time complexity.
4. **Run spectral clustering for longer**, to minimize the total loss.