Ali Bahreinian

Project Title:

Manipulating Graph Neural Networks

IE 7945 Project Course

Supervisor: Prof. Radhakrishnan

Table Of Content:

Section 1: Introduction to the Project

Background:

The rapid advancements in machine learning and data science have revolutionized various fields, enabling more accurate predictions, insightful data analysis, and efficient decision-making processes. Graph-based data structures have gained prominence due to their ability to represent complex relationships and interactions in data. Applications of graph-based models span numerous domains, including social network analysis, biological network modeling, recommendation systems, and many more.

In the context of this project, we focus on the application of machine learning models on graph-structured data. The primary objective is to explore and implement advanced graph neural network (GNN) techniques to enhance the understanding and prediction capabilities of such models. By leveraging recent developments in GNNs (Graph Neural Networks), including Graph Convolutional Networks (GCNs), Graph Attention Networks (GATs), and GraphSAGE, we aim to improve the performance of machine learning tasks on graph data.

Objectives:

The main objectives of this project are:

1. Literature Review: Conduct a comprehensive review of key research papers related to machine learning models for graph-structured data, focusing on data preprocessing techniques, feature extraction, and relevant applications.
2. Dataset Identification and Acquisition: Identify and obtain suitable datasets for graph-based machine learning tasks. Ensure the datasets are diverse and representative of real-world scenarios.
3. Model Implementation and Evaluation: Implement state-of-the-art GNN models and evaluate their performance on the selected datasets.
4. Analysis and Reporting: Analyze the results, document the findings, and provide insights into the effectiveness of different GNN techniques.

Significance:

This project's significance lies in its potential to advance the understanding and application of GNNs in various domains. By systematically reviewing the literature, implementing, and evaluating innovative models, this project contributes to the body of knowledge in machine learning and data science. Moreover, the findings can be applied to real-world problems, offering improved solutions for complex data analysis tasks.

Scope:

The scope of this project includes:

Literature Review: Reviewing at least 5-7 key research papers from recent years and seminal works in GNNs.

- Dataset Acquisition: Identifying 2-3 relevant datasets from reputable sources like the UCI Machine Learning Repository, Kaggle, and SNAP.
- Model Implementation: Implementing GCN, GAT, and GraphSAGE models using popular machine learning frameworks.
- Evaluation: Comparing the performance of different models on the selected datasets using standard metrics.
- Documentation: Preparing a comprehensive report detailing the methodology, results, and conclusions.

Section 2: Annotated Bibliography

1. **Hamilton, W. L., Ying, R., & Leskovec, J. (2017). Inductive Representation Learning on Large Graphs.**

   Summary: This paper introduces GraphSAGE, a framework designed to generate low-dimensional node embeddings for large graphs using node features and neighborhood sampling techniques. GraphSAGE focuses on inductive learning, enabling the model to generalize to unseen nodes by learning a function that generates embeddings based on the features and structure of local neighborhoods.

   Key Insights: The authors propose different aggregator functions (mean, LSTM, pooling) to aggregate node features from local neighborhoods. The framework is evaluated on node classification tasks using citation and Reddit post data, demonstrating significant improvements over existing methods.

   Relevance: GraphSAGE is particularly relevant for projects involving dynamic or evolving graphs where new nodes frequently appear, such as social networks, recommendation systems, and biological networks.

2. **Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2018). Graph Attention Networks.**

   Summary: This paper presents Graph Attention Networks (GATs), which leverage self-attention mechanisms to assign different importances to nodes in a neighborhood, thereby addressing some limitations of previous graph convolutional methods. GATs can be applied to both inductive and transductive learning tasks.

Key Insights: By using masked self-attention layers, GATs can efficiently compute node features while attending over their neighbors. This approach allows for varying neighborhood sizes and improves computational efficiency. The authors demonstrate state-of-the-art performance on several benchmark datasets, including citation networks and protein-protein interaction networks.

Relevance: GATs are ideal for projects that need to handle graph data with variable neighborhood sizes and require efficient computation of node importance, such as social network analysis and biological network modeling.

3. **Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. (2017). Neural Message Passing for Quantum Chemistry.**

   Summary: This paper introduces a general framework for message-passing neural networks (MPNNs) applied to quantum chemistry. The framework effectively predicts molecular properties by aggregating information from neighboring atoms in a molecule.

   Key Insights: The authors use message passing to aggregate information iteratively from neighboring nodes (atoms), applying this technique to predict quantum properties of molecules. MPNNs demonstrate superior performance in predicting molecular properties compared to traditional methods.

   Relevance: MPNNs are significant for projects in computational chemistry or materials science where accurate predictions of molecular properties are essential for drug discovery and material design.

4. **Kipf, T. N., & Welling, M. (2017). Semi-Supervised Classification with Graph Convolutional Networks.**

   Summary: This seminal paper introduces Graph Convolutional Networks (GCNs) for semi-supervised learning on graph-structured data, achieving state-of-the-art performance on citation network datasets. GCNs perform convolution operations directly on graphs by aggregating feature information from a node’s local neighborhood.

   Key Insights: The layer-wise propagation rule based on the graph Laplacian enables GCNs to capture the structure of graph data effectively. The authors show that GCNs can significantly improve classification accuracy on citation networks while being computationally efficient.

   Relevance: GCNs are foundational for projects involving semi-supervised learning on graph data, such as citation networks, social networks, or biological networks, where labeled data is scarce but unlabeled data is abundant.

5. **Lee, J., Lee, I., & Kang, J. (2019). Self-Attention Graph Pooling.**

Summary: This paper proposes SAGPool, a novel graph pooling method that uses self-attention to select the most informative nodes. This approach improves the performance of graph neural networks by effectively reducing the graph size while preserving important structural information.

Key Insights: SAGPool integrates self-attention mechanisms to identify and retain the most relevant nodes in a graph, addressing the challenge of handling large-scale graphs. The authors demonstrate the effectiveness of SAGPool on various graph classification tasks.

Relevance: SAGPool is valuable for projects requiring efficient graph pooling techniques to handle large-scale graphs, such as hierarchical graph representation learning and large-scale graph classification.

6. **Zitnik, M., & Leskovec, J. (2017). Predicting Multicellular Function through Multi-Layer Tissue Networks.**

   Summary: This paper presents a method for predicting multicellular function using multi-layer tissue networks. The approach integrates data from different tissue-specific networks to predict the function of genes and proteins across different biological contexts.

   Key Insights: The authors use a multi-layer network model to capture the interactions between genes and proteins across different tissues. This method improves the accuracy of functional predictions by leveraging complementary information from multiple biological networks.

   Relevance: This method is highly relevant for projects in systems biology and bioinformatics, where understanding the functional roles of genes and proteins in different tissues is crucial for disease research and drug development.

Section 3: Dataset Overview and Repository Details

1. **Dataset: Cora Citation Network**

   Source: Available from various academic repositories (e.g., https://linqs.soe.ucsc.edu/data)

   Dataset Size: 2,708 nodes, 5,429 edges

   Data Attributes: Node features (bag-of-words representation of document contents), class labels (research paper topics)

   Availability & Quality: Publicly available, widely used benchmark dataset, high-quality curated data

   Challenges: Sparse connectivity, imbalanced class distribution

2. **Dataset: Reddit Post Data**

Source: Available on Kaggle and other data repositories

Dataset Size: 232,965 posts, 11,663,972 edges

Data Attributes: Node features (word embeddings of post content), labels (subreddit categories)

Availability & Quality: Publicly available, large-scale dataset with rich information, requires preprocessing

Challenges: High computational complexity, dynamic nature of data

3. Dataset: Protein-Protein Interaction (PPI)

Source: SNAP repository (Stanford Network Analysis Project)

Dataset Size: 56,944 nodes, 818,716 edges

Data Attributes: Node features (biological attributes), labels (gene ontology terms)

Availability & Quality: Publicly available, high-quality dataset for biological network analysis

Challenges: Multi-label classification, high-dimensional feature space

References:

1. W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," presented at the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 2017.

2. P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," presented at the International Conference on Learning Representations (ICLR 2018), 2018.

3. J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," presented at the 34th International Conference on Machine Learning (ICML 2017), Sydney, NSW, Australia, 2017.

4. T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," presented at the International Conference on Learning Representations (ICLR 2017), 2017.

5. J. Lee, I. Lee, and J. Kang, "Self-attention graph pooling," presented at the 36th International Conference on Machine Learning (ICML 2019), Long Beach, CA, USA, 2019.

6. M. Zitnik and J. Leskovec, "Predicting multicellular function through multi-layer tissue networks," Bioinformatics, vol. 33, no. 14, pp. 190-198, 2017.