H-UniMP: A Heterogeneity-Aware Unified Message Passing Model for Citation Networks

Advanced Machine Learning - CS4681

Progress Evaluation

Index: 210110B

Name: De Zoysa P.V.K.

Contents

List of Figures

1. Project Planning

1.1 Objective

The main objective of this project is to enhance the Unified Message Passing (UniMP) model so that it can be applied effectively to heterogeneous citation networks. While UniMP has shown state-of-the-art performance on homogeneous graphs, it does not explicitly consider multiple node and edge types, which are fundamental in real-world citation networks. For example, in the DBLP-Citation-Network V12, the graph contains papers, authors, and venues as nodes, with relations such as cites, writes, and published in. Traditional UniMP treats all nodes and edges as the same type, which leads to a loss of rich semantic information. This limitation creates a research gap: how can we unify feature and label propagation in a framework that also respects heterogeneity and semantic relations? This project aims to solve this problem by developing Heterogeneity-Aware UniMP (H-UniMP), which extends UniMP with relation-aware attention. By doing so, the model can capture both structural and semantic information in citation graphs, leading to more accurate and robust node classification.

1.2 Baseline

UniMP (Shi et al., 2021): Designed for homogeneous graphs which unifies feature and label propagation via masked label prediction.

1.3 Enhancement

The main enhancement in this project is to adapt the Unified Message Passing (UniMP) model for heterogeneous citation networks by introducing relation-aware mechanisms and improving the training strategy.

- Extend UniMP to handle multiple node and edge types, enabling it to operate effectively on heterogeneous graphs such as the DBLP-Citation-Network V12.
- Introduce relation-aware attention, so that different types of edges (e.g., cites vs. writes) are weighted differently during message passing, allowing the model to capture the

semantics of heterogeneous relations.

- Optimize masked label prediction hyperparameters, particularly the masking rate, through systematic tuning. This ensures robust training and improved performance in semi-supervised node classification.

1.4 Dataset

DBLP-Citation-Network V12 (Heterogeneous citation benchmark)

- Nodes - papers, authors, venues
- Relations - cites, writes, published in
- Task - Node classification (e.g., predicting research field/topic of papers)

1.5 Deliverables

- H-UniMP Implementation - A complete codebase extending UniMP with relation-aware attention for heterogeneous citation networks.
- Comparative Results - Benchmarking H-UniMP against UniMP, R-GCN, and HAN on DBLP V12.
- Ablation Studies - Experiments to evaluate the impact of relation-aware attention and the effect of different masking rates in the training strategy.
- Research Paper - 6 to 8 pages conference-style paper with methodology, results, and analysis.
- Reproducible Repository - Public GitHub repository with well-documented code, datasets, and instructions.

2. Literature Review

Graphs are a natural way to represent data in many domains, such as citation networks, social networks, biological interactions, and recommendation systems. In these graphs, the task of node classification is very important, where we try to predict labels of nodes (like classifying papers into topics in a citation network) when only part of the nodes are labeled. Since labeling large datasets is expensive, researchers focus on semi-supervised learning, where the model learns from a small number of labeled nodes and generalizes to many unlabeled ones.

One of the most influential methods for this problem is Graph Neural Networks (GNNs). GNNs work by repeatedly aggregating features from neighboring nodes, a process known as message passing. Popular GNN variants include the Graph Convolutional Network (GCN), GraphSAGE, and Graph Attention Networks (GAT). These models propagate and combine node features through the graph structure, making them strong at capturing both node attributes and graph connectivity. However, GNNs mainly focus on feature propagation and often ignore the explicit use of label information during inference.

On the other hand, Label Propagation Algorithms (LPA) take a different approach. Instead of relying on node features, LPA spreads the known labels of some nodes across the graph based on the assumption that connected nodes are likely to share the same label. While simple and effective, LPA cannot make use of node features and becomes less powerful when the graph structure alone is not enough to make correct predictions.

Since GNNs and LPA are both message passing algorithms but focus on different aspects (features vs. labels), many researchers have tried to combine the strengths of both. For example, APPNP (Gasteiger et al. (2018)) [1] predicts soft labels using a GNN and then propagates them with Personalized PageRank, and GCN-LPA (Wang & Leskovec, 2019) [2] uses LPA as a regularization term during training. These approaches improved results to some extent, but they still could not fully unify feature propagation and label propagation in a single framework for both training and inference.

To address this gap, Shi et al. (2021) [3] proposed the Unified Message Passing model (UniMP). UniMP is a graph transformer-based model that takes both node features and observed labels as inputs. A key innovation is its use of masked label prediction, inspired by the masked word prediction strategy in BERT. During training, some labels are randomly hidden (masked), and the model is tasked with predicting them. This prevents the problem of label leakage (overfitting to already known labels) and encourages the model to learn to propagate label information effectively to unlabeled nodes. By embedding labels into the same space as features and processing both with a graph transformer, UniMP successfully unifies feature propagation and label propagation.

Experimental results on the Open Graph Benchmark (OGB) datasets showed that UniMP achieved state-of-the-art performance. On ogbn-products, it reached 82.56% accuracy, on ogbn-proteins it achieved 86.42% ROC-AUC, and on ogbn-arxiv it achieved 73.11% accuracy, outperforming several strong baselines like DeeperGCN and GCNII. Further ablation studies demonstrated that incorporating label information significantly improved results compared to models that only used features.

In summary, the literature shows that while GNNs and LPAs individually perform well, their integration is more powerful. UniMP represents a breakthrough by directly combining feature and label propagation in a unified framework with masked label prediction, setting a strong baseline for future research in semi-supervised node classification. However, UniMP has mainly been applied to homogeneous graphs. Extending it to heterogeneous graphs, such as citation networks with multiple node and edge types, remains an open research direction. This is the area where our work aims to contribute.

3. Methodology

3.1 Baseline Reproduction

The first step of this project is to reproduce the original UniMP model on a homogeneous dataset such as OGBN-Arxiv. This is necessary to establish a reliable baseline for comparison and to confirm that our implementation is correct. By running UniMP on a single-type graph, I can observe its performance under the same conditions described in the original paper. This step also highlights the limitation of UniMP in ignoring node and edge heterogeneity, which motivates our proposed extension.

3.2 Dataset Preparation

After establishing the baseline, I move to the DBLP-Citation-Network V12, which is a large-scale heterogeneous citation network. This dataset contains multiple node types, including papers, authors, and venues, as well as multiple relation types such as cites (paper to paper), writes (author to paper), and published in (paper to venue). The first task is to preprocess this dataset into a format that can be used for graph learning. This includes,

- Converting categorical attributes into vectorized features
- Constructing adjacency lists for each relation type
- Splitting labeled nodes into training, validation, and test sets

3.3 H-UniMP Design

I extend UniMP into Heterogeneity-Aware UniMP (H-UniMP) by introducing a relation-aware transformer as the enhancement.

- **Relation-Aware Graph Transformer**
  In UniMP, all edges are treated equally when propagating information between nodes. In H-UniMP, I assign relation-specific attention weights so that different types of edges,

such as cites versus writes, are modeled separately. This makes the propagation process sensitive to the meaning of each relation type.

3.4 Training Strategy

To address the problem of label leakage, where the model might overfit to already observed labels, I adopt a masked label prediction strategy. During training, a fixed proportion of the available labels are randomly masked, and the model is required to predict them using the remaining labels and node features. For example, a certain percentage of paper labels may be hidden in each batch, forcing the model to rely on its propagation mechanism rather than memorizing labels directly.

The training objective combines two parts,

- Cross-entropy loss, which supervises the classification of labeled nodes
- Masked label prediction loss, which encourages the model to recover the masked labels

To better understand the effect of this strategy, I vary the masking rate across different experiments (e.g., 20%, 40%, 60%) as part of sensitivity analysis. This helps identify the most effective configuration and ensures that the model generalizes well, simulating real-world conditions where many labels in a citation network are unavailable.

3.5 Evaluation

I evaluate H-UniMP on the DBLP-Citation-Network V12 benchmark and compare it against three strong baselines,

- Original UniMP (homogeneous baseline)
- R-GCN (relation-aware GNN)
- HAN (heterogeneous attention network with meta-paths)

The main metrics I use are,

- Node classification accuracy (overall correctness)
- Macro-F1 score (balanced performance across classes)

- Micro-F1 score (performance across all nodes, weighted by class size)

3.6 Analysis and Ablation Studies

To evaluate the contribution of the proposed enhancement, I perform ablation studies focusing on the effect of relation-aware attention. Specifically, I compare the performance of the baseline UniMP model with and without relation-aware attention to measure how much this modification improves node classification in heterogeneous citation networks.

In addition to this, I carry out a sensitivity analysis on the masked label prediction strategy. By varying the label masking rate during training (for example, testing 20%, 40%, and 60%), I analyze how the availability of observed labels affects the overall model performance.

These experiments help identify whether relation-aware attention provides significant benefits, and how different masking configurations influence learning stability and accuracy. Together, the ablation and sensitivity studies ensure that the effectiveness of the proposed enhancements is well understood.

4. Project Timeline

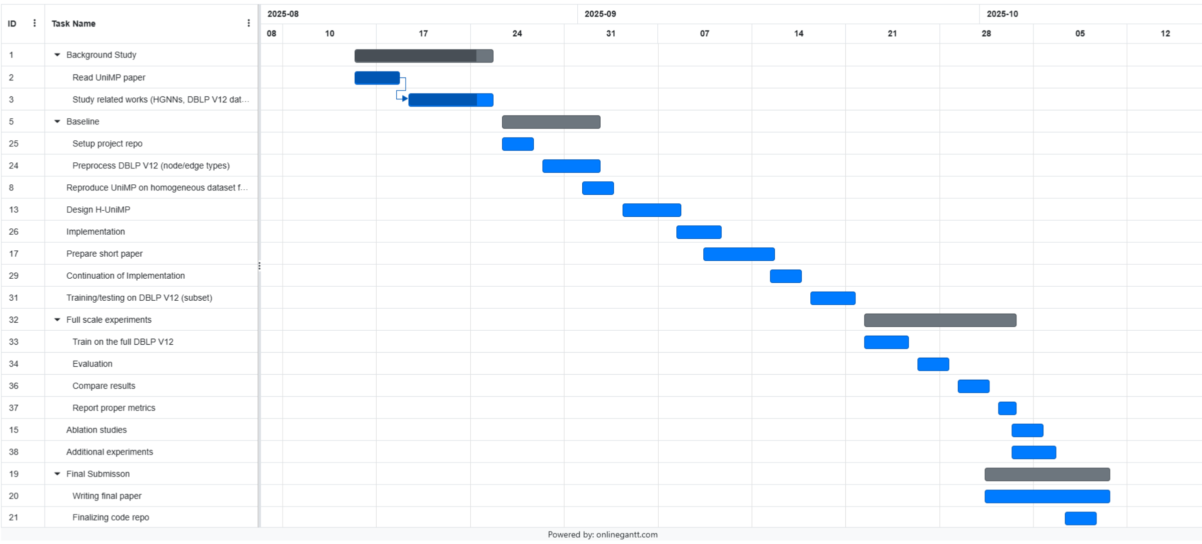

Figure 4.1: Gantt chart illustrating the project schedule.

Bibliography

[1] Gasteiger, J., Bojchevski, A., & Günnemann, S. (2018, October 14). *Predict then Propagate: Graph Neural Networks meet Personalized PageRank*. arXiv.org., 2018, 598–603. https://arxiv.org/abs/1810.05997https://arxiv.org/abs/1810.05997

[2] Wang, H., & Leskovec, J. (2020, February 17). *Unifying graph convolutional neural networks and label propagation.* https://arxiv.org/abs/2002.06755https://arxiv.org/abs/2002.06755

[3] Shi, Y., Huang, Z., Feng, S., Zhong, H., Wang, W., & Sun, Y. (2020, September 8). *Masked Label Prediction: Unified Message Passing Model for Semi-Supervised Classification.* arXiv.org. https://arxiv.org/abs/2009.03509https://arxiv.org/abs/2009.03509