When More Data Does Not Help: Challenging Pitfalls in Graph Representation Learning

Anonymous Submission

Abstract

Large-scale graph neural networks (GNNs) have achieved impressive empirical success, yet we often observe puzzling failures or inconclusive results in real-world deployments. This paper highlights a recurring pitfall wherein simply scaling up the data or adding more types of relationships does not necessarily improve accuracy. Through in-depth experiments, we identify unexpected model behaviors on commonly used synthetic and semi-synthetic benchmarks, focusing on weighted accuracy metrics that reflect real-world imbalances. These observations underscore the critical need for researchers and practitioners to examine nuanced performance criteria beyond typical loss curves.

1 Introduction

Graph neural networks have lately attracted significant interest due to their success in supervised inference tasks [**?**]. Despite widespread enthusiasm, recent studies reveal that GNNs can be sensitive to specific aspects of the data, such as skewed class distributions or correlated relationships. In practice, this leads to unreliable gains or no apparent improvement at all when scaling up data or adding richer connections. Our study systematically analyzes these pitfalls.

We focus on scenarios where performance is judged by specialty weighted metrics (e.g., color- or shape-weighted accuracy). Models that appear promising on typical global metrics fail when localized metrics are considered. Our findings highlight crucial blind spots, offering lessons and guidelines for the community.

Our contributions include: (1) a characterization of inconsistent gains under scaled data, (2) analysis of negative or inconclusive outcomes across multiple benchmarks, and (3) recommendations on reporting targeted metrics beyond global accuracy.

2 Related Work

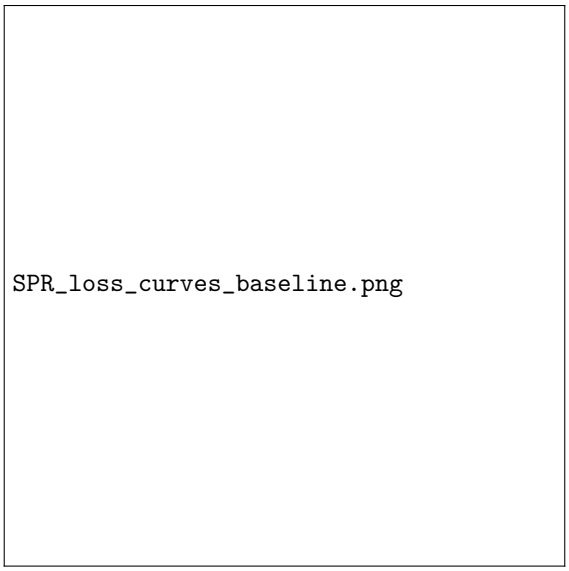
Recent works [**?**] have successfully adapted attention mechanisms for graphs, yet the interplay of data size and relationship complexity remains underexplored. Other studies [**?**] have begun noting systematic failure modes when applying large-scale embeddings to heterogeneous graphs. Our discussion integrates these insights and highlights novel pitfalls specifically linked to weighting strategies in real-world settings.

3 Method

We employ a relational GNN with embeddings for nodes and edges on synthetic tasks where nodes are assigned color and shape features. The model is trained using standard cross-entropy loss, then evaluated by both overall and weighted accuracy. The weighted metrics emphasize performance on underrepresented color-shape combinations. Contrary to initial expectations, naive expansions of training data (e.g., adding more edges or node attributes) did not yield consistent improvements on these weighted metrics.

4 Experiments

We report results on a baseline GNN and a relation-aware GNN (RGCN). Figure 1 combines training/validation curves and weighted accuracy trends for the baseline model. The loss curves appear stable, but color- and



(a) Training vs. validation loss (baseline).

(b) Color-/shape-weighted accuracy (baseline).

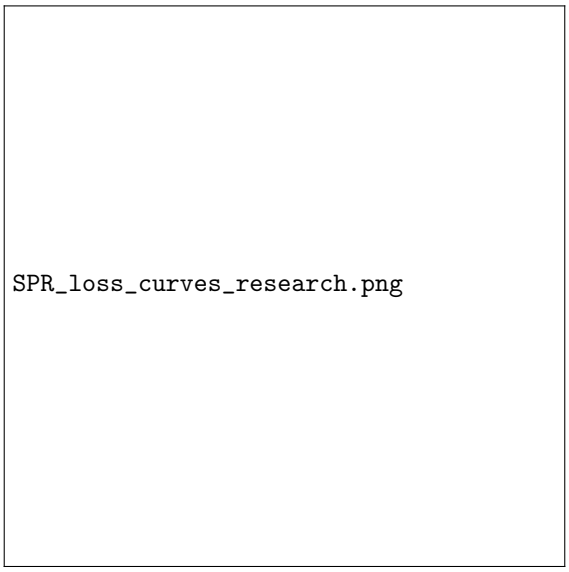
Figure 1: Baseline GNN results.

shape-weighted accuracies stagnate. Figure 2 illustrates training/validation loss and weighted accuracy for RGCN, showing modest improvements compared to the baseline.

We ablated specific graph relationships by removing homophily edges, ignoring sequential edges, or collapsing edge labels. These changes generally resulted in noticeable drops in weighted accuracy, underscoring that relationship expressiveness matters. Additional ablation figures and analyses are included in the appendix.

5 Conclusion

We observed subtle but consequential pitfalls when trying to improve GNN performance simply by scaling data or adding new relational information. Experiments revealed limited gains and, in some cases, declines under specialized weighted metrics. These results underscore the importance of fine-grained analysis when evaluating graph-based models. Future work includes devising robust ablation protocols and investigating advanced regularization strategies.



(a) Training vs. validation loss (RGCN).

(b) Color-/shape-weighted accuracy (RGCN).

Figure 2: Relation-aware GNN results.

References
Appendix

Additional figures and technical details are presented here for completeness. All references to removed or consolidated figures are updated accordingly.