Rethinking Deeper Graph Neural Networks: Negative Results and Surprising Overfitting Issues

Anonymous Submission
ICBINB Workshop @ ICLR 2025

Abstract

We investigate deeper Graph Convolutional Networks (GCNs) and Relational GCNs (R-GCNs) on synthetic datasets. Despite standard regularization and well-tuned hyperparameters, our results show that deeper models can dramatically overfit, leading to marginal or no real-world performance gains. We highlight reasons why these deeper networks may fail to generalize, suggesting future directions for robust graph learning.

1 Introduction

Deep Graph Neural Networks (GNNs) offer expressive capacity for complex relational tasks. However, reports of deeper GNN success have been inconsistent, and many real-world deployments remain shallow [? ?]. In this paper, we explore multiple attempts to build deeper GCN and R-GCN architectures. Our results reveal frustrating overfitting on seemingly straightforward synthetic tasks. We highlight these pitfalls, why they matter for real-world scenarios, and how negative or partial results can guide both future architecture design and rigorous benchmarking.

2 Related Work

Recent research suggests that training deeper GNNs is nontrivial due to oversmoothing, exploding gradients, and other issues [? ?]. Performance improvements often come only with carefully crafted layer-wise strategies or specialized skip connections. Nevertheless, many works do not systematically report negative or inconclusive results. By documenting our failed or partially successful attempts, we expand on existing evidence that GNN depth is not a panacea.

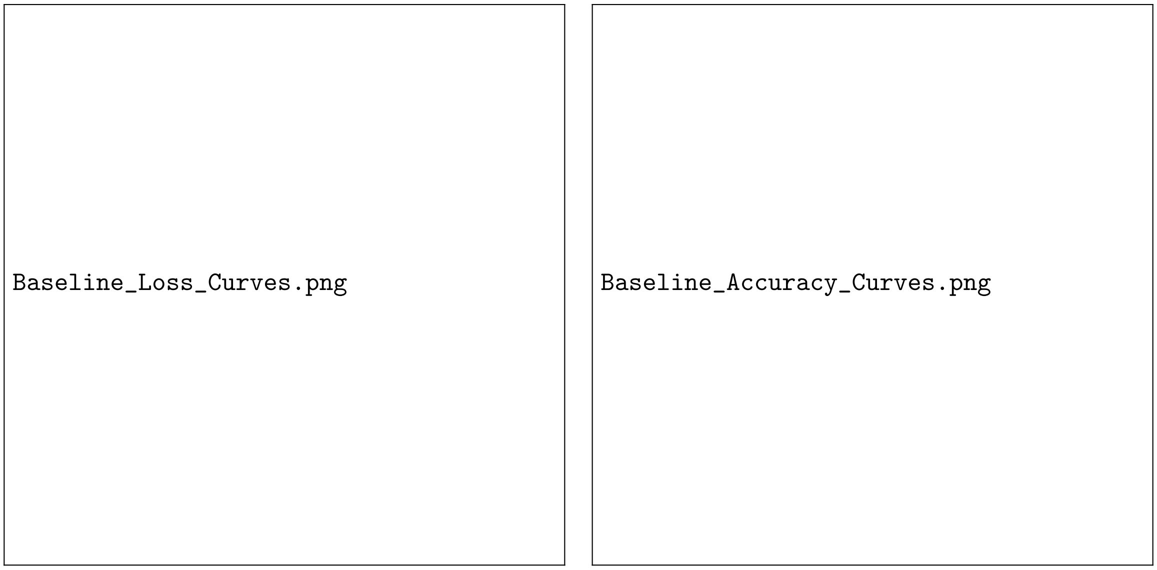
3 Method / Problem Discussion

We experiment with standard GCN and R-GCN designs on various synthetic graphs that feature entity relations and sequential edges. Our baseline is a 2-layer GCN using a cross-entropy loss. We further apply an R-GCN with additional relation coefficients. Our objective was to ascertain if deeper models—with 6 or more layers—improve generalization when modeling complex relational structures.

4 Experiments

We first present baseline performance in Figure 1. The GCN shows increasing accuracy early in training, but we observe overfitting after several epochs. Adding deeper layers exacerbates this trend.

Figure 2 depicts representative R-GCN results. Although the model can capture richer relational patterns, it also suffers from heavy overfitting and does not conclusively outperform the shallower baseline on final test metrics.



(a) Training and validation loss

(b) Accuracy trends over epochs

Figure 1: Baseline GCN results. Overfitting arises within the first 20 epochs, with the validation loss diverging while training performance continues to improve.

Although each architecture trains smoothly for a time, we do not see robust generalization. Neither additional layers nor added relational features yield consistent improvements. We believe that further research on specialized regularization or architectural constraints is needed to fully unlock deeper GNN benefits.

5 Conclusion

We detailed experiments showing how deeper GCNs and R-GCNs can fail to generalize, offering warnings to practitioners looking to deploy large-scale GNNs. Future work may involve new normalization strategies, domain-specific pretraining, or stronger regularization. Our findings underscore that deeper does not necessarily mean better and that more open reporting of negative results can benefit the community.

(a) Training vs. validation loss

(b) Test metrics summary (CWA, SWA, etc.)

Figure 2: Deeper R-GCN attempts. We observe intensified overfitting and modest or no real gains in terms of final test set performance.

References

A Additional Experiments and Visualizations

We place supplementary results here for completeness. Some tasks do benefit slightly from deeper representations, but the overall improvements remain inconsistent. Figure 3 shows performance on synthetic expansions. Figures 4, 5, and 6 illustrate ablation studies and additional diagnostic plots that did not fit in the main paper.

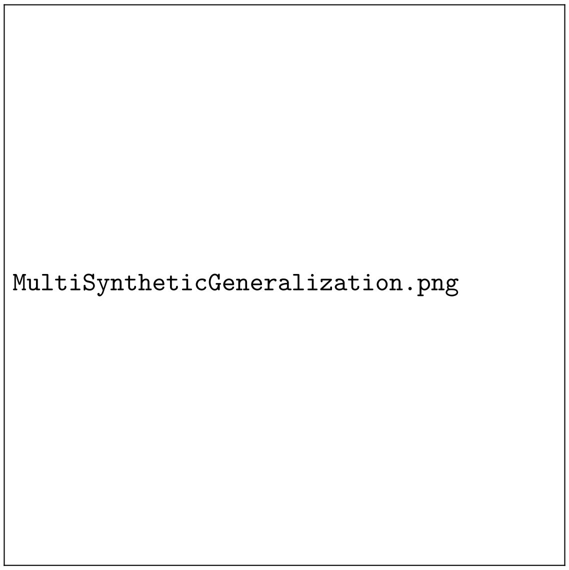


Figure 3: Performance across multiple synthetic expansions.

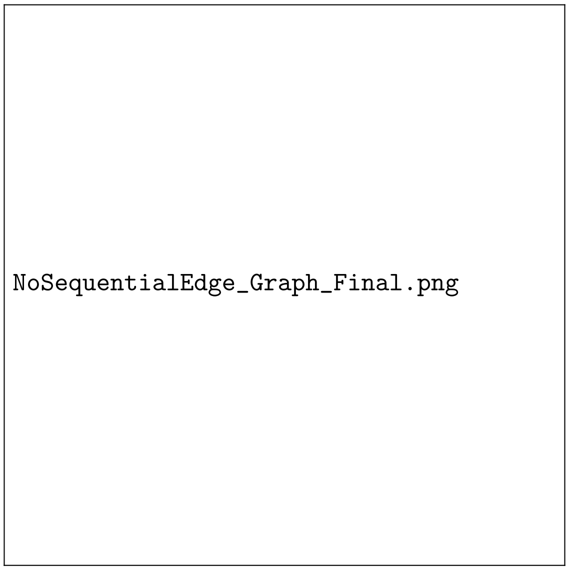


Figure 4: Removing sequential edges.

SPR_RGCN_MultiMetric_Curves.png

Figure 5: Detailed R-GCN performance curves across multiple metrics.

ShallowGNN_Ablation.png

Figure 6: Very shallow GNN ablations underfit, highlighting depth importance.