LEVERAGING GRAPH NEURAL NETWORKS FOR ENHANCED SYNTHETIC POLYRULE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

We explore Graph Neural Networks (GNNs) for the Synthetic PolyRule Reasoning (SPR) task, which involves classification of symbolic sequences governed by poly-factor rules. Existing sequence-based models may overlook latent relational structures within these sequences. We represent each sequence as a graph, connecting tokens according to position, shape, and color attributes. We find that while training accuracy improves, the model struggles to generalize and ultimately fails to surpass reported state-of-the-art metrics on color-weighted (CWA) or shape-weighted (SWA) accuracy. Our empirical findings reveal potential overfitting, especially as complexity-weighted performance on validation data remains volatile, suggesting that more specialized designs or regularization strategies are needed to handle the relational complexity of SPR.

1 INTRODUCTION

The Synthetic PolyRule Reasoning (SPR) task involves identifying patterns in symbolic sequences governed by multiple interacting factors. Unlike simple rule-based sequence tasks, SPR data often contains relationships that go beyond purely sequential dependencies, making it challenging to discover hidden correlations among tokens. Sequence-based models (e.g., LSTMs (Hochreiter & Schmidhuber, 1997) or Transformers (Vaswani et al., 2017)) primarily focus on positional order. However, we hypothesize that more explicit modeling of relational and structural attributes will yield stronger representations. We investigate Graph Neural Networks (GNNs), which handle node- and edge-level relationships more naturally (Hamilton et al., 2017; Meng et al., 2024). Our work focuses on converting sequences into graph structures, aiming to capture these latent relationships more effectively.

We report that our GNN-based architecture achieves steadily increasing training accuracies but shows limited validation improvement on SPR benchmarks, indicating overfitting and inconclusive gains over established methods. These observations highlight pitfalls in applying GNNs to tasks with intricate factor interactions and underscore the importance of better regularization and data handling.

2 RELATED WORK

Numerous approaches leverage sequential models for symbolic reasoning, including RNN variants and Transformer-based architectures (Goodfellow et al., 2016; Vaswani et al., 2017; Hochreiter & Schmidhuber, 1997), but these primarily process data in a strictly linear fashion. GNNs have been explored in domains involving structural and relational data (Hamilton et al., 2017; Meng et al., 2024), including systematic reasoning tasks (Khalid & Schockaert, 2024). Moreover, specialized benchmarks have been proposed in related fields (Teney et al., 2019), yet few systematically focus on symbolic poly-factor sequences like SPR. Our study builds on these lines, casting the SPR sequences into graphs to investigate whether relational modeling can exceed conventional sequence-based methods.

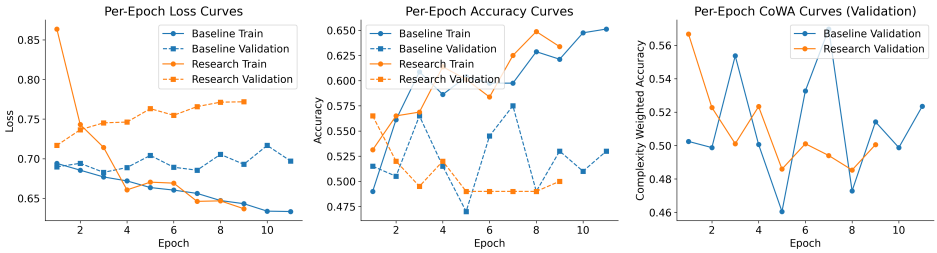

Figure 1: Aggregated training and validation curves for GNN-based and baseline approaches (left: loss, middle: accuracy, right: complexity-weighted accuracy). The validation metrics remain volatile, indicating difficulty in capturing the data's relational complexity.

3 Method

We represent each sequence in SPR as a graph where each token becomes a node. Edges connect (1) consecutive tokens and (2) tokens sharing the same shape or color. Node features encode shape, color, and position. A GNN then aggregates local neighborhood information through graph convolution layers (Hamilton et al., 2017), followed by pooling for sequence-level classification. Unlike fully sequential architectures, which may ignore cross-token similarities, this graph-based approach explicitly captures structural relationships. We implement variants of GraphSAGE (Hamilton et al., 2017), comparing different ways of adding edges and regularizers.

4 Experiments

We train and validate on the `SPR_BENCH` dataset, partitioned into train, development, and test. We evaluate performance on standard accuracy, complexity-weighted accuracy, and examine color-weighted (CWA) or shape-weighted (SWA) metrics. Despite strong training performance (e.g., training accuracies above 0.63), validation and test metrics fluctuate, with no decisive improvement relative to sequence-based baselines. Figure 1 shows an aggregated plot of training and validation curves for our GNN variants and a baseline. These results suggest that while GNNs learn patterns in training data, they risk overfitting, particularly on sequences containing many confounding factors. Furthermore, complexity-weighted or color/shape-weighted metrics do not reveal a clear performance advantage over simpler baselines.

We also explore ablations removing positional embeddings or batch normalization, observing only marginal differences on test metrics. Detailed confusion matrices for these ablations are shown in Appendix A.1, indicating patterns of misclassification that remain largely unchanged without these design elements.

5 Conclusion

We investigated GNN-based approaches for SPR, motivated by the hope of leveraging structural relationships for improved classification. Our results demonstrate that while GNNs learn effectively on training data, they struggle to maintain consistency on validation sets, leading to minimal net improvements and failure to surpass the reported SOTA on color- and shape-weighted metrics. These observations highlight a common pitfall: relational modeling can be powerful but also prone to overfitting in complex symbolic tasks. Future research should investigate advanced regularization, improved graph construction strategies, or specialized attention mechanisms to help GNNs better generalize in poly-factor rule settings.

References

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

William L. Hamilton, Z. Ying, and J. Leskovec. Inductive representation learning on large graphs. *ArXiv*, abs/1706.02216, 2017.

Sepp Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9:1735–1780, 1997.

Irtaza Khalid and S. Schockaert. Systematic relational reasoning with epistemic graph neural networks. 2024.

Lei Meng, Zhonglin Ye, Yanlin Yang, and Haixing Zhao. Deepmcgcn: Multi-channel deep graph neural networks. *Int. J. Comput. Intell. Syst.*, 17:41, 2024.

Damien Teney, Peng Wang, Jiewei Cao, Lingqiao Liu, Chunhua Shen, and A. Hengel. V-prom: A benchmark for visual reasoning using visual progressive matrices. pp. 12071–12078, 2019.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and I. Polosukhin. Attention is all you need. pp. 5998–6008, 2017.

SUPPLEMENTARY MATERIAL

A ADDITIONAL FIGURES AND DETAILS

A.1 CONFUSION MATRICES

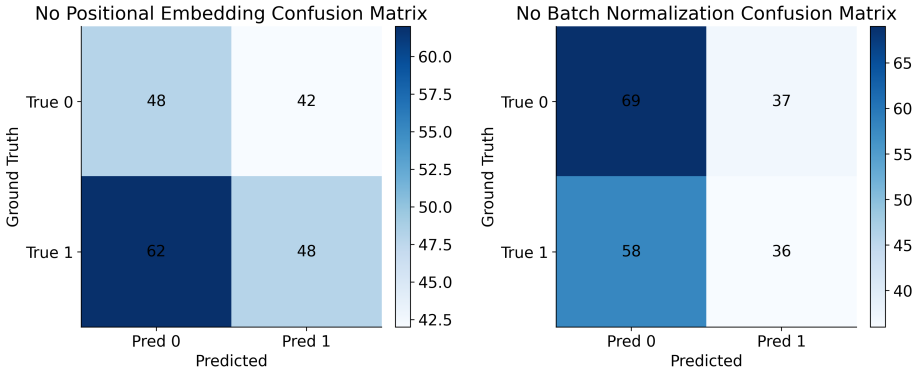

Figure 2: Confusion matrices for "No Positional Embedding" (left) and "No Batch Normalization" (right) ablations. Despite changes, the distribution of errors remains similar, suggesting that these architectural choices are not the sole factor limiting generalization.