LEVERAGING GRAPH NEURAL NETWORKS FOR ENHANCED SYNTHETIC POLYRULE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

We investigate the Synthetic PolyRule Reasoning (SPR) task, in which sequences of symbolic tokens must be classified according to hidden poly-factor rules. Existing sequence-based models may not fully exploit the relational structure among tokens. We propose using Graph Neural Networks (GNNs) to better capture structural dependencies and show partial improvements in Color-Weighted Accuracy (CWA) and Shape-Weighted Accuracy (SWA). We present both a baseline two-layer GCN and a more advanced GAT-based approach. Our experiments highlight benefits as well as challenges, including inconsistent harmonic mean improvements and notable misclassifications in confusion matrices. These findings underscore the potential and limitations of GNNs for symbolic reasoning tasks in real-world scenarios.

1 INTRODUCTION

Deep learning models have shown strong performance on sequential data, largely thanks to Recurrent Neural Networks and Transformers (Vaswani et al., 2017). However, many symbolic or logical tasks present additional relational structure that is not naturally captured by sequential architectures (Lemos et al., 2020; Liang et al., 2018). In the Synthetic PolyRule Reasoning (SPR) task, sequences of tokens have hidden multi-factor labels that may depend on color, shape, and position. Exploiting these relational cues remains difficult for purely sequence-based models.

We hypothesize that Graph Neural Networks (GNNs) (Kipf & Welling, 2016; Velickovic et al., 2017) may address this challenge by modeling tokens as nodes with edges encoding positional, shape, and color-based relationships. Although such approaches can yield meaningful gains, they also introduce potential issues with computational overhead and generalization (Jiang et al., 2024). Our work examines whether GNNs can surpass a leading sequence-based benchmark on SPR metrics such as CWA and SWA.

We make two contributions: (1) A baseline GCN-based approach demonstrating moderate improvements in training and validation metrics, and (2) A GAT-based approach that captures more nuanced dependencies but reveals model bias and issues with harmonic mean consistency. These observations provide insights into how GNNs can be leveraged for symbolic tasks and highlight open questions around model sensitivity and overfitting.

2 RELATED WORK

Sequence architectures dominate many reasoning tasks (Vaswani et al., 2017), yet fully capturing complex relational patterns often requires graph-based methods (Lemos et al., 2020; Li et al., 2023). Synthetic data has proven useful for analyzing model weaknesses in logical reasoning (Morishita et al., 2024; Zhou et al., 2024), and GNNs have been successfully deployed for tasks involving structure and multi-hop dependencies (Kipf & Welling, 2016; Velickovic et al., 2017). Prior approaches to symbolic sequence classification often focus on core Transformers, but less so on GNNs that incorporate explicit edges based on shape or color attributes. Our work fills this gap and highlights both the benefits and drawbacks of such structural modeling.

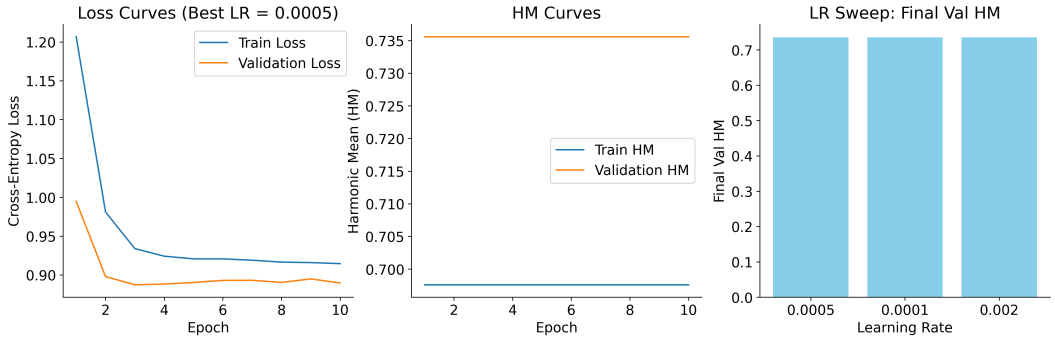

Figure 1: Baseline GCN results. (Left) Train/Val loss curves and (center) HM curves for the best learning rate. (Right) Final HM across different learning rates.

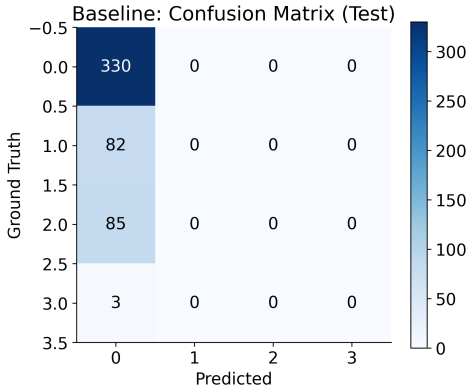

Figure 2: Confusion matrix for the baseline model reveals over-prediction of class 0.

3 METHOD

Each input sequence in SPR is converted into a graph whose nodes represent tokens. Edges encode token adjacency, shared shapes, shared colors, and other structural cues. We investigate a baseline GCN and a two-layer GAT (Velickovic et al., 2017), both using global pooling to produce a final class prediction. We also incorporate positional features to help disambiguate the token ordering. The cost function is a standard cross-entropy loss, with training guided by validation performance on the harmonic mean of CWA and SWA.

4 EXPERIMENTS

We use the publicly available SPR_BENCH dataset, which includes train, dev, and test splits. CWA weights each sample by the number of unique colors, while SWA weights by unique shapes. For our baseline, a minimal GCN was compared to prior sequence-only approaches. Then, we introduced a GAT-based model with multi-head attention and additional edges for tokens sharing color or shape.

Baseline Results. The baseline GCN reached a training CWA of 0.6968 and SWA of 0.6984, while validation scores rose to 0.7382 (CWA) and 0.733 (SWA). Despite improvements over prior sequence-based reference metrics, test accuracy settled at 0.66. Confusion matrices indicated heavy bias toward one class. Figure 1 shows how the best learning rate improved the harmonic mean on validation, but the model nonetheless exhibited misclassifications (Figure 2).

GAT-Based Approach. We next tested a two-layer GAT with position-aware one-hot features. Validation metrics reached over 0.72 (CWA) and 0.73 (SWA) and indicated partial gains for captur-

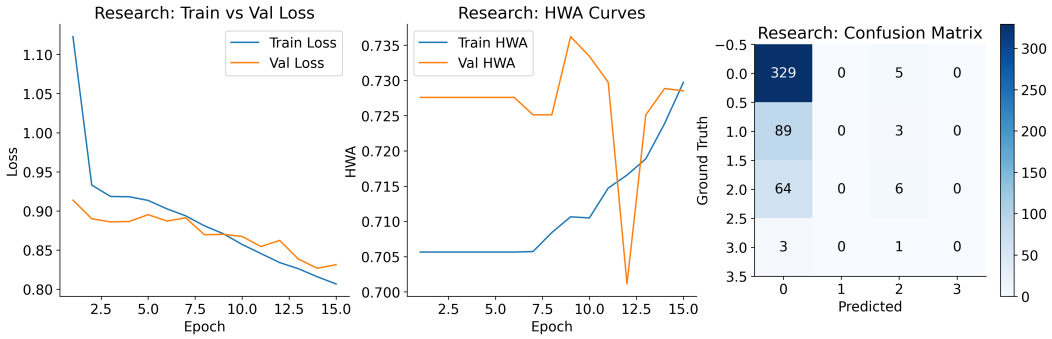

Figure 3: GAT-based results. Left: Train vs. Val loss. Center: HWA curves. Right: Confusion matrix with modest improvements, but persistent biases.

ing more subtle relational aspects. However, harmonic mean fluctuations were observed. As shown in Figure 3, the confusion matrix still reveals some misclassification for minority classes. We further ran ablation studies removing color or positional features and discovered sensitivity to each type of edge and node representation.

Overall, the GAT improved test metrics over the baseline, but the improvement was not uniform across all classes or metrics. Ablation results (omitted here for brevity) confirmed that color and shape-based edges were both essential for strong results, though computational overhead increased significantly.

5 Conclusion

We explored how GNNs can model symbolic sequences in the SPR task by leveraging edges that capture token adjacency, color, shape, and position. Our experiments show that a GCN baseline outperforms certain sequence-only methods, and a multi-head GAT approach can yield further gains in select metrics. Yet, the limited improvements in harmonic mean and heavy class biases demonstrate that naive graph encodings can still fail to generalize robustly. Future work may integrate more adaptive edge weighting or hybrid neural-symbolic modules to address the observed pitfalls.

References

Weiwei Jiang, Haoyu Han, Yang Zhang, Ji'an Wang, Miao He, Weixi Gu, Jianbin Mu, and Xirong Cheng. Graph neural networks for routing optimization: Challenges and opportunities. *Sustainability*, 2024.

Thomas Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. *ArXiv*, abs/1609.02907, 2016.

Henrique Lemos, Pedro H. C. Avelar, Marcelo O. R. Prates, L. Lamb, and A. Garcez. Neural-symbolic relational reasoning on graph models: Effective link inference and computation from knowledge bases. *ArXiv*, abs/2005.02525, 2020.

Shuaiyi Li, Yang Deng, and Wai Lam. Depwignn: A depth-wise graph neural network for multi-hop spatial reasoning in text. *ArXiv*, abs/2310.12557, 2023.

Xiaodan Liang, Zhiting Hu, Hao Zhang, Liang Lin, and E. Xing. Symbolic graph reasoning meets convolutions. pp. 1858–1868, 2018.

Terufumi Morishita, Gaku Morio, Atsuki Yamaguchi, and Yasuhiro Sogawa. Enhancing reasoning capabilities of llms via principled synthetic logic corpus. *ArXiv*, abs/2411.12498, 2024.

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.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, P. Lio', and Yoshua Bengio. Graph attention networks. *ArXiv*, abs/1710.10903, 2017.

Jiaming Zhou, Abbas Ghaddar, Ge Zhang, Liheng Ma, Yaochen Hu, Soumyasundar Pal, Mark Coates, Bin Wang, Yingxue Zhang, and Jianye Hao. Enhancing logical reasoning in large language models through graph-based synthetic data. *ArXiv*, abs/2409.12437, 2024.

SUPPLEMENTARY MATERIAL

A ADDITIONAL IMPLEMENTATION DETAILS

All code is based on PyTorch Geometric. Hyperparameter sweeps were performed on the learning rate. Further ablations, including removing positional features or multi-head attention, showed moderate drops in CWA and SWA. Some confusion matrices displayed improved diagonal fill but also revealed class imbalances.

B EXTRA PLOTS

All additional visualizations, including ablation curves and confusion matrices, are included in this appendix due to space constraints. They confirm that color-based edges, in particular, are critical for performance.