LEVERAGING GRAPH NEURAL NETWORKS FOR ENHANCED SYNTHETIC POLYRULE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

In Synthetic PolyRule Reasoning (SPR), sequences of symbolic tokens must be classified according to hidden poly-factor rules. Existing sequence-based models often struggle to capture relational and structural aspects underlying these tasks, hindering real-world deployment when symbolic reasoning is required. We propose using Graph Neural Networks (GNNs) to represent sequences as graphs, with edges encoding relations such as position, color, or shape. Experiments with GraphSAGE and an RGCN variant indicate that GNNs can improve color-focused metrics while exhibiting overfitting pitfalls. Our findings highlight both benefits and challenges of graph-based reasoning approaches for SPR.

1 INTRODUCTION

The Synthetic PolyRule Reasoning (SPR) task classifies sequences of colored shapes according to multiple hidden rules. Standard sequence models (e.g., RNNs or Transformers) can ignore global structural dependencies that matter in real-world scenarios demanding explicit relational reasoning (Barbiero et al., 2023; Himabindu et al., 2023). In industrial and biomedical domains, for instance, missing relational cues can lead to inaccurate predictions that undermine practical trust (Kipf & Welling, 2016; Khalid & Schockaert, 2024).

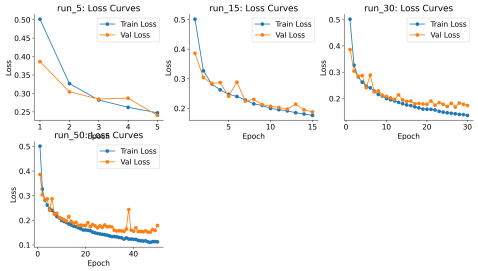
In this paper, we leverage Graph Neural Networks (GNNs) to encode both local and non-adjacent connections in SPR. Our main contributions include: (1) a graph-based encoding of colored shape sequences, (2) experiments that reveal partial accuracy gains on color-focused metrics, and (3) a discussion of pitfalls such as overfitting. While validation performance soars above 0.95 on weighted metrics, we observe notable test drops, illuminating how domain mismatch and model complexity affect generalization.

2 RELATED WORK

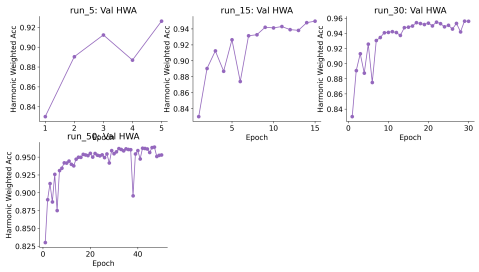
Graph-based learning is now a cornerstone of deep relational modeling (Hamilton et al., 2017; Kipf & Welling, 2016; Liao et al., 2022). In symbolic reasoning domains, neural-symbolic fusion can yield interpretability gains (Barbiero et al., 2023; Himabindu et al., 2023; Özgür Yılmaz et al., 2016). Weighted metrics (e.g., shape- or color-based) have been used to counter the accuracy paradox in imbalanced tasks (Uddin, 2019). However, scaling GNNs and transferring performance from validation to deployment remains challenging (Khemani et al., 2024).

3 METHOD

We transform each SPR sequence into a node-labeled graph that incorporates shape, color, and position. Edges reflect properties such as adjacency, shared color, or shape. Our baseline applies GraphSAGE (Hamilton et al., 2017) to learn node embeddings, pooled into sequence predictions. To test enhanced relational modeling, we modify edges to incorporate separate relation types in an RGCN. We measure Color-Weighted Accuracy (CWA), Shape-Weighted Accuracy (SWA), and a combined Harmonic Weighted Accuracy (HWA), acknowledging the complexity of multi-factor signals.

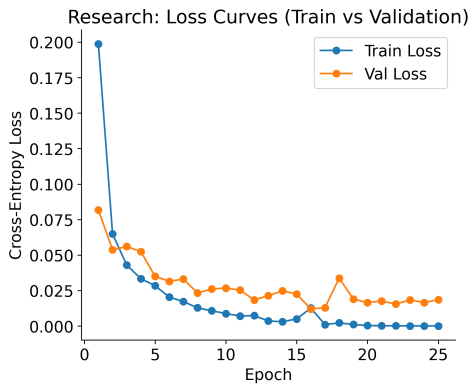


(a) Baseline Loss Curves

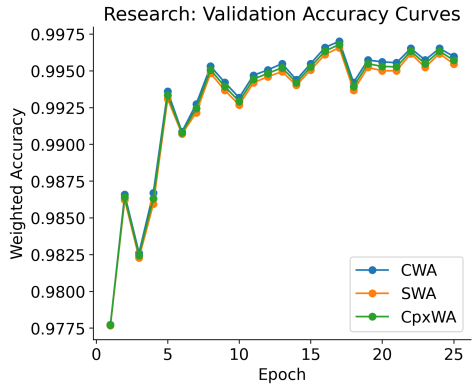


(b) Baseline Val HWA

Figure 1: Performance of the GraphSAGE baseline. (a) Training and validation loss converge. (b) Validation HWA stabilizes after about 30 epochs, but test metrics drop.



(a) Research Loss Curves



(b) Research Val Metrics

Figure 2: RGCN-based approach. (a) Train/validation loss curves continue to decline. (b) Weighted accuracies (CWA, SWA, CpxWA) indicate high validation performance, with color-based metrics peaking around epoch 20.

Hyperparameters. We use the Adam optimizer with an initial learning rate of $1\mathrm{e}-3$, batch size 32, and L2 weight decay of 5×10^{-4}. Early stopping monitors validation HWA.

4 Experiments

Setup. We train on SPR_BENCH with train/dev/test splits. Data includes sequences of up to 30 tokens, each with shape and color. We keep 70% of data for training, 10% for validation, and 20% for test.

Baseline Results. As shown in Figure 1, our GraphSAGE model achieves validation HWA of 0.96–0.97 (averaged across runs). By epoch 30, we see stable validation performance in subfigure (b). However, test results degrade to CWA=0.682, SWA=0.638, and HWA=0.659, illustrating significant overfitting.

RGCN Variant. We further incorporate relation-specific edges (same color, same shape, positional neighbors) using an RGCN. Figure 2 displays training and validation curves. Validation metrics exceed 0.99 for color-based measurements (subfigure b), but test metrics (CWA=0.701, SWA=0.653) remain lower than the validation results. These findings suggest GNNs capture domain-specific cues well in training but struggle with generalization.

Real-World Pitfalls. Our observations reflect a common pitfall: high validation scores can mask overfitting when the domain distribution diverges from test data. Industrial applications requiring robust shape recognition may see performance degrade if training data lacks shape variance.

5 Conclusion

We investigated GNN-based modeling for SPR, revealing partial improvements in color-focused accuracy but persistent gaps to test performance. Notably, shape-based misclassifications and domain mismatch pose threats to real-world deployment. Future work could involve augmented data generation, more sophisticated regularization, or domain adaptation strategies. Our findings highlight both the promise and pitfalls of GNNs for symbolic tasks involving structural relationships.

References

Pietro Barbiero, Gabriele Ciravegna, Francesco Giannini, Mateo Espinosa Zarlenga, Lucie Charlotte Magister, A. Tonda, Pietro Lio', F. Precioso, M. Jamnik, and G. Marra. Interpretable neural-symbolic concept reasoning. *ArXiv*, abs/2304.14068, 2023.

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

Modi Himabindu, Revathi V, Manish Gupta, Ajay Rana, Pradeep Kumar Chandra, and H. S. Abdulaali. Neuro-symbolic ai: Integrating symbolic reasoning with deep learning. In *IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics Engineering*, volume 10, pp. 1587–1592, 2023.

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

Bharti Khemani, S. Patil, K. Kotecha, and Sudeep Tanwar. A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions. *Journal of Big Data*, 11:1–43, 2024.

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

Ningyi Liao, Dingheng Mo, Siqiang Luo, Xiang Li, and Pengcheng Yin. Scara: Scalable graph neural networks with feature-oriented optimization. *ArXiv*, abs/2207.09179, 2022.

M. Uddin. Addressing accuracy paradox using enhanched weighted performance metric in machine learning. *2019 Sixth HCT Information Technology Trends (ITT)*, pp. 319–324, 2019.

Özgür Yılmaz, A. Garcez, and Daniel L. Silver. A proposal for common dataset in neural-symbolic reasoning studies. 2016.

Supplementary Material

A Additional Hyperparameter Details

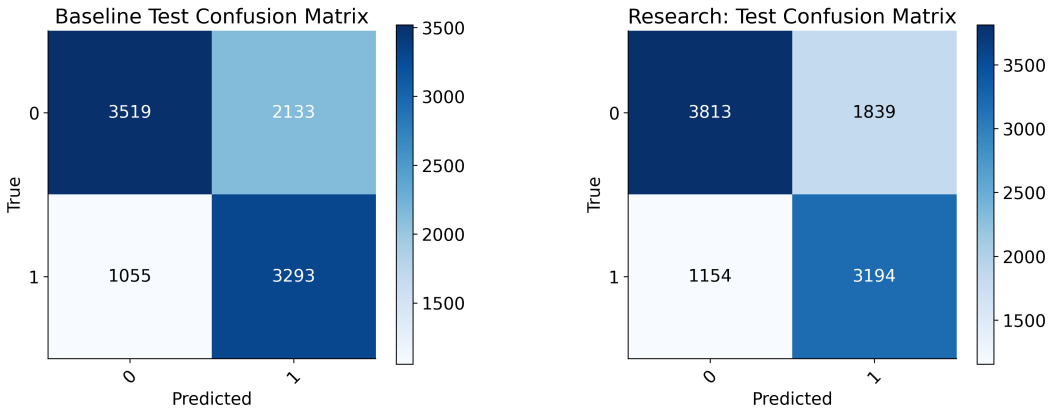
We used up to 64 hidden units in each GraphSAGE or RGCN layer. Dropout of 0.2 was applied before the final readout. Batch normalization was used to stabilize training. In all cases, we used a maximum of 100 epochs with early stopping on validation HWA.

B Additional Figures

B.1 Confusion Matrices

C Further Discussion

Analysis suggests that color edges provide strong local signals, while shape edges introduce subtle dependencies. The test domain possibly contains unusual shape composition that leads to out-of-distribution generalization failures. Balancing color- and shape-based features remains an open direction for robust real-world deployment.



(a) Baseline Confusion Matrix (b) Research Confusion Matrix

Figure 3: Confusion matrices for shape predictions, illustrating misclassifications that primarily occur in subsets of visually or positionally similar shapes.

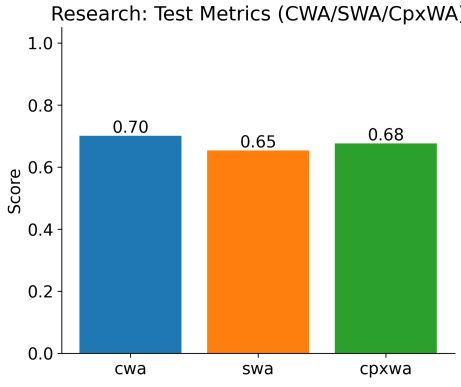


Figure 4: Test metrics for the research (RGCN) model over multiple runs, showing variable CWA and SWA across seeds but consistent patterns of shape confusion.