UNVEILING HIDDEN PATTERNS: SYMBOLIC GLYPH CLUSTERING FOR ENHANCED POLYRULE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Symbolic Pattern Recognition (SPR) requires models to interpret complex rules from abstract glyph sequences. We investigate whether clustering symbolic glyphs via latent features can help discover these hidden patterns, thereby improving SPR accuracy. Specifically, we propose to apply clustering on glyphs before training a reasoning model on the Synthetic PolyRule Reasoning (SPR) task. Experiments on the `SPR_BENCH` dataset show that, while symbolic clustering can reveal latent structures, substantial gaps arise between development and test performances. This emphasizes the need for more robust methods that generalize beyond local data distributions in symbolic reasoning.

1 INTRODUCTION

Synthetic PolyRule Reasoning (SPR) involves deciphering latent rules governing symbolic glyph sequences and predicting sequence labels based on color and shape (Minervini et al., 2020; Sileo, 2024). Despite progress in deep symbolic learning, real deployment often experiences unpredictable generalization gaps. Our work focuses on a clustering-based approach to group symbolic glyphs, followed by training a sequence model with these cluster IDs. We ask whether glyph clustering can illuminate latent rules and improve performance.

In practical scenarios, bridging the gap between promising validation scores and often lower test performance is crucial (Li et al., 2024). We analyze internal pitfalls: do we see actual improvements or just overfitting to dev data? Our contributions include: (a) a pipeline for clustering glyphs and transforming sequences into cluster histograms or sequences; (b) an empirical investigation of the dev–test gap to highlight known pitfalls; (c) negative or inconclusive results underscoring the challenges of robust SPR.

2 RELATED WORK

Neuro-symbolic approaches have tackled logical reasoning with learned rules, for instance in theorem proving (Minervini et al., 2020). Symbolic learning methods have used clusters primarily for visual patterns (Oliveira & Marçal, 2023), though Snell et al. (2017) demonstrate effectiveness in few-shot classification. For abstract symbolic tasks, symbolic feature extraction has been typically overshadowed by end-to-end neural methods. Our method emphasizes silhouette-based validation (Semoglou et al., 2025; Sheng et al., 2025) and the potential for partial improvement. Furthermore, a domain gap arises between dev and test sets, reminiscent of domain shift observed in tasks like cross-domain feature extraction (Guan & Zhang, 2023).

3 METHOD

We extract glyph-level latent representations (e.g., via a pre-trained model akin to BERT embeddings (Guan & Zhang, 2023)), then apply clustering (e.g., K-Means or DBSCAN). Cluster assignments augment or replace raw glyph tokens in an SPR classifier. We employ silhouette scores to verify clustering quality (Oliveira & Marçal, 2023; Semoglou et al., 2025). Our classifier uses these cluster features either as histograms or as input embeddings (GRU-based). This approach aims to capture underlying shapes and colors in symbolic space.

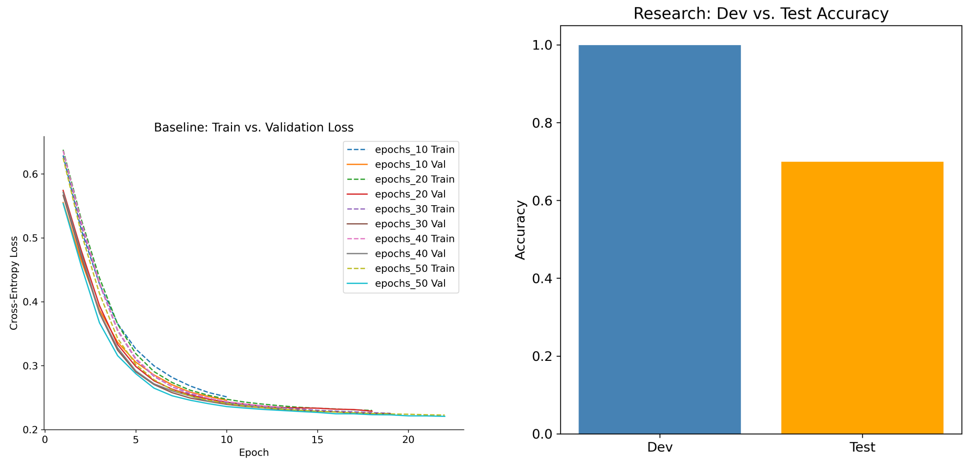
We use the `SPR_BENCH` data, introduced by Sileo (2024), splitting into train (20k), dev (5k), and test (10k). The labels reflect color or shape parity. The metrics are Color-Weighted Accuracy (CWA) and Shape-Weighted Accuracy (SWA). We also report a derived Harmonic Color-Shape Accuracy (HCSA) or extended metrics (SNWA) for some experiments.

4 Experiments

We summarize two representative experiments here. Full details, codes, and additional ablations appear in the appendix. Across these experiments, we observe strong dev performance but a notable drop on the test set.

Baseline vs. Tuning. As shown in Figure 1(a), a baseline model with a grid search over epochs sees good dev HCSA rapidly plateauing near 0.92, but the test HCSA remains around 0.64. This discrepancy demonstrates potential overfitting or data shifts.

Glyph Clustering + GRU. In Figure 1(b), we embed clustered glyph IDs into a bidirectional GRU. The approach achieves near-perfect dev metrics (CWA/SWA ≈ 1.0), yet test accuracy is roughly 0.70. While clustering clarifies latent structures, the lesson learned is that easy dev success does not always translate to test robustness.



(a) Baseline Loss Curves

(b) Research Dev vs. Test

Figure 1: **(a)** Baseline approach: training vs. validation loss across different epoch budgets. **(b)** Large dev–test gap for the clustering-based variant.

We further tested ablations like removing bidirectionality or ignoring glyph clustering. These occasionally reduce complexity but still exhibit dev–test divergence.

5 Conclusion

Symbolic glyph clustering for SPR can expose latent feature patterns and boost dev scores; however, consistently high test performance remains elusive. Our findings, including dev–test discrepancies, confirm that naive adoption of clustering does not guarantee generalizable improvements. Future research should seek techniques to mitigate overfitting, possibly by diversifying training data or incorporating domain adaptation strategies.

References

Kunlun Guan and Yangsen Zhang. Cross domain few shot ner via data enhancement invariant domain feature extraction based on diffusion model. *2023 IEEE Intl Conf on Parallel Distributed*

Processing with Applications, Big Data Cloud Computing, Sustainable Computing Communications, Social Computing Networking (ISPA/BDCloud/SocialCom/SustainCom), pp. 980–988, 2023.

Zi'ang Li, Xiaofeng Jia, Jie Zhang, Peimin Zhu, Wei Cai, Hao Zhang, and Zhiying Liao. Characterizing near-surface velocity structures via deep learning. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024.

Pasquale Minervini, Sebastian Riedel, Pontus Stenetorp, Edward Grefenstette, and Tim Rocktäschel. Learning reasoning strategies in end-to-end differentiable proving. pp. 6938–6949, 2020.

Mafalda Oliveira and A. Marçal. Clustering lidar data with k-means and dbscan. pp. 822–831, 2023.

Aggelos Semoglou, Aristidis Likas, and John Pavlopoulos. Silhouette-guided instance-weighted k-means. *ArXiv*, abs/2506.12878, 2025.

ChenNingZhi Sheng, Rafal Kustra, and Davide Chicco. Comparative analysis of unsupervised clustering techniques using validation metrics: Study on cognitive features from the canadian longitudinal study on aging (clsa). *ArXiv*, abs/2504.12270, 2025.

Damien Sileo. Scaling synthetic logical reasoning datasets with context-sensitive declarative grammars. *ArXiv*, abs/2406.11035, 2024.

Jake Snell, Kevin Swersky, and R. Zemel. Prototypical networks for few-shot learning. pp. 4077–4087, 2017.

SUPPLEMENTARY MATERIAL

A ADDITIONAL IMPLEMENTATION AND ABLATIONS

All code, datasets, and further plots (e.g., confusion matrices, more ablation curves) are provided in the repository. Figures include: `NoBidirectionalGRU_Loss_Curves.png`, `MeanPoolingEncoder_Loss_Curves.png`, `NoSequencePacking_Loss_Curves.png`, and related metric charts.

We implemented our clusters using `sklearn`'s `KMeans` and validated them with silhouette scores (Semoglou et al., 2025; Sheng et al., 2025). Hyperparameters such as the number of clusters (8–16) and the embedding dimension were scanned. Yet, partial improvements on dev do not reliably translate into test gains, highlighting required future work on distributional robustness.