UNVEILING HIDDEN PATTERNS: SYMBOLIC GLYPH CLUSTERING FOR ENHANCED POLYRULE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Symbolic Pattern Recognition (SPR) presents a unique challenge for machine learning models, requiring them to decipher hidden rules in abstract symbol sequences. We hypothesize that clustering symbolic glyphs before rule extraction can reveal important cues and improve performance. We evaluate this hypothesis on the SPR_BENCH dataset (Özgür Yılmaz et al., 2016; Xie et al., 2025), focusing on Color-Weighted Accuracy (CWA) and Shape-Weighted Accuracy (SWA). Our results highlight partial success, especially on shape-based metrics, while color-based performance remains below the anticipated threshold. This underscores the need for deeper analysis of clustering approaches for symbolic reasoning.

1 INTRODUCTION

Deep learning methods achieve impressive results in fields like image recognition and language modeling (Goodfellow et al., 2016), yet tasks requiring symbolic reasoning often remain challenging. Within Synthetic PolyRule Reasoning (SPR), purely data-driven models may fail to accurately capture abstract rules. This paper explores the idea of glyph-level clustering: turning each symbol into a cluster-derived representation, with the goal of enhancing interpretability and improving performance.

Our contributions include: (1) A glyph clustering approach that yields low-dimensional symbolic abstractions. (2) A detailed empirical investigation, exposing both the benefits and the pitfalls of clustering-based symbolic reasoning. (3) Experiments demonstrating promising shape-based accuracy improvements, although color-based metrics remain below the desired target.

2 RELATED WORK

Neuro-symbolic research (Daniele et al., 2022) merges trainable neural mechanisms with symbolic abstractions to achieve better interpretability. Clustering methods like k-means (Hartigan & Wong, 1979) and metric-based prototypes (Snell et al., 2017) can help discover latent groupings, but the risk of poor clustering in high-dimensional spaces is well-documented (Kukreti, 2021). Dimensionality reduction (Pomerantsev, 2014; Hasan & Abdulazeez, 2021) further streamlines clustering by eliminating redundant features. Despite these advances, systematic investigations of clustering pitfalls in symbolic rule reasoning remain sparse.

3 METHOD / PROBLEM DISCUSSION

We propose a pipeline that first extracts numeric embeddings from symbolic glyphs, then applies k-means to form a small vocabulary of discrete clusters. The resulting cluster indices form the input to a rule predictor, which may be a recurrent or transformer-based classifier. The hypothesis is that learned clusters expose latent regularities that facilitate rule discovery.

Risk Factors and Limitations. Unstable clustering can degrade performance. Large-scale data introduce computational overhead for repeated clustering. Moreover, domain shifts might render glyph groupings inadequate. These pitfalls, while potentially limiting, can reveal paths to more robust neuro-symbolic systems.

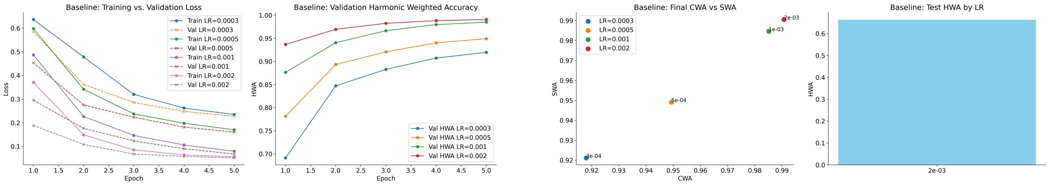

Figure 1: **Baseline Experiments.** **(Left)** Two subplots show training/validation loss and validation HWA across different learning rates over epochs. **(Right)** Scatter plot of final color vs. shape accuracy, and a bar chart for test HWA under one chosen learning rate.

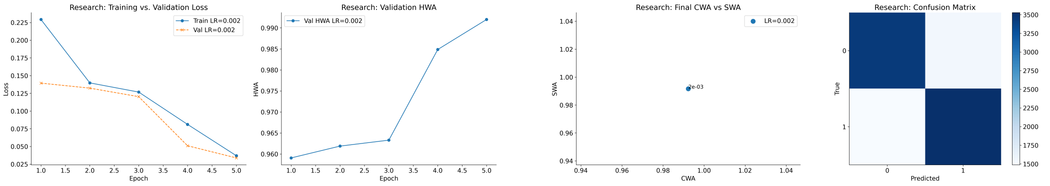

Figure 2: **Transformer Experiments.** **(Left)** Training/validation loss and validation HWA across epochs. **(Right)** Final color vs. shape accuracy in a scatter plot and a confusion matrix revealing predominant diagonal entries but some notable misclassifications.

4 Experiments

We use SPR_BENCH, featuring train/dev/test splits with unknown rules. Each sequence is labeled according to shape and color relationships. We measure Color-Weighted Accuracy (CWA) and Shape-Weighted Accuracy (SWA). Our baseline GRU model, with 8–16 clusters, achieves a test CWA of 63.3% and an SWA of 69.6%. A transformer model marginally improves CWA to 63.4% and SWA to 69.7%. Although shape-based performance surpasses the reference of 65.0% SWA, the color-based accuracy remains below the 70.0% goal.

In Figure 1, the left panel displays how varying learning rates affects training and validation loss, as well as validation HWA. Notably, some learning rates converge faster but can plateau at suboptimal accuracies. The right panel includes a scatter plot of final color vs. shape accuracies and a bar chart highlighting test HWA for a specific learning rate. These metrics illustrate the trade-offs: shape accuracy is more robust, while color accuracy lags.

Figure 2 shows the transformer-based approach. Although we observe slightly improved SWA, color-based tasks remain problematic, indicating possible missed patterns in glyph clustering. Confusion matrices reveal that most predictions are correct, but certain symbol combinations lead to errors that reduce color accuracy.

5 Conclusion

Our study highlights both the subtle gains and notable pitfalls of glyph clustering for symbolic rule reasoning. While improved shape-based performance suggests that structural embeddings can be beneficial, failing to capture color-related attributes underscores the need for refined clustering strategies. Future work may explore adaptive clustering or specialized head networks to better handle color-based distinctions.

References

Alessandro Daniele, Tommaso Campari, Sagar Malhotra, and L. Serafini. Deep symbolic learning: Discovering symbols and rules from perceptions. *ArXiv*, abs/2208.11561, 2022.

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

J. Hartigan and M. A. Wong. A k-means clustering algorithm. 1979.

Basna Mohammed Salih Hasan and A. Abdulazeez. A review of principal component analysis algorithm for dimensionality reduction. volume 2, 2021.

Anil Kukreti. A survey of clustering algorithms for high-dimensional data mining. *Mathematical Statistician and Engineering Applications*, 2021.

Alexey L. Pomerantsev. Principal component analysis (pca). *Encyclopedia of Autism Spectrum Disorders*, 2014.

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

Zhuohan Xie, Dhruv Sahnan, Debopriyo Banerjee, Georgi N. Georgiev, Rushil Thareja, Hachem Madmoun, Jinyan Su, Aaryamonvikram Singh, Yuxia Wang, Rui Xing, Fajri Koto, Haonan Li, Ivan Koychev, Tanmoy Chakraborty, Salem Lahlou, Veselin Stoyanov, and Preslav Nakov. Finchain: A symbolic benchmark for verifiable chain-of-thought financial reasoning. *ArXiv*, abs/2506.02515, 2025.

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

Supplementary Material

A Implementation Details

Our implementation uses PyTorch. Symbolic glyphs are mapped to simple numeric ASCII-based embeddings, optionally reduced via PCA (Pomerantsev, 2014; Hasan & Abdulazeez, 2021) before k-means clustering (Hartigan & Wong, 1979). We vary k from 8 to 32, trading off expressiveness and over-partitioning. After determining cluster assignments, each sequence is represented by cluster indices. We train models for up to 20 epochs using Adam with an initial learning rate in {1e-3, 2e-3, 3e-3} and a batch size of 64. We repeat all experiments across three random seeds to mitigate chance variations.

B Ablation and Extended Results

To examine the impact of clustering, we tested architectures with no clustering (direct symbolic embeddings), random partitioning, and alternative positional encodings. Figures below show that no or random clustering yields volatile performance, while omitting positional encodings slightly degrades final results. Nonetheless, color-based metrics remain lower than shape-based variants.

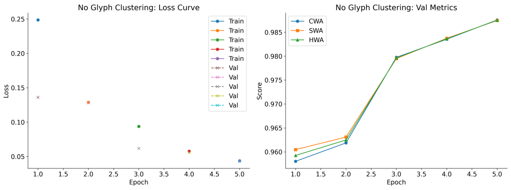

Figure 3: **No Glyph Clustering.** Training/validation loss curves and validation metrics are more volatile.

B.1 CLS Token Pooling Variant

We also experimented with a CLS-token pooling approach in transformers, instead of average pooling (see Figure 6). Training converged faster, yet final test metrics remained nearly the same.

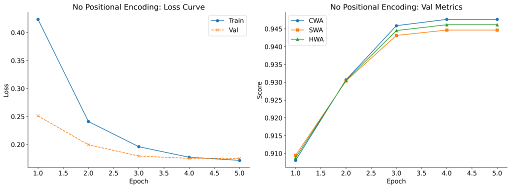

Figure 4: **No Positional Encoding.** Training remains stable, though final accuracies diminish slightly.

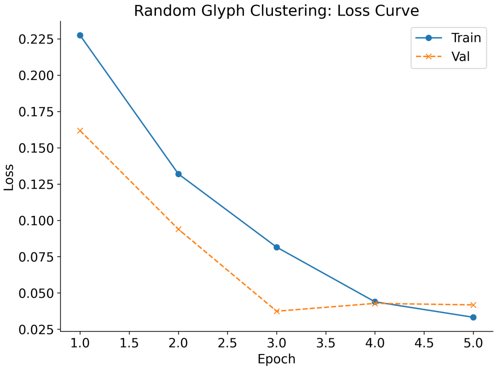

Figure 5: **Random Glyph Clustering.** Performance consistently lags behind the baseline clustering approach.

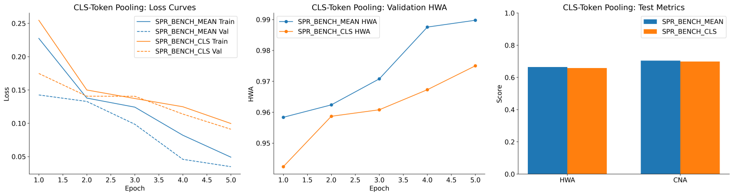

Figure 6: **CLS vs. Mean Pooling.** Both approaches yield similar final results, with minor convergence differences.