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 uncover rules governing sequences of abstract glyphs. We explore clustering glyphs via auto-encoder embeddings and show that cluster labels alongside token embeddings can boost Weighted Accuracy from about 0.04 to near-perfect levels on SPR_BENCH. However, these inflated scores may not generalize when glyphs deviate from training distributions, suggesting caution in relying on clustering alone.

1 INTRODUCTION

Machine learning systems often struggle to extrapolate from small or abstract symbol sets, where latent attributes drive inferences (Goodfellow et al., 2016; Devlin et al., 2019; Mondorf & Plank, 2024). We investigate Symbolic PolyRule Reasoning (SPR), where shape-color glyphs form sequences mapped to labels. Traditional token embeddings sometimes climb from near 0.04 Weighted Accuracy toward 0.9 or higher, indicating partial capability. We study whether clustering symbolic glyphs prior to inference can improve performance. Our analysis suggests that near-perfect metrics may arise from memorizing ephemeral patterns, underscoring the gap between in-distribution accuracy and real-world robustness (Alotaibi et al., 2024; Yu et al., 2024).

2 RELATED WORK

Clustering has long aided pattern discovery in data (Hartigan & Wong, 1979; Sreedhar et al., 2017; Deng, 2020), while auto-encoders capture compressed representations (Lee et al., 2023). Hybrid symbolic approaches often rely on embeddings (Snell et al., 2017), but few systematically cluster glyphs. Our method merges these concepts to gauge if grouping tokens by latent similarity clarifies symbolic tasks.

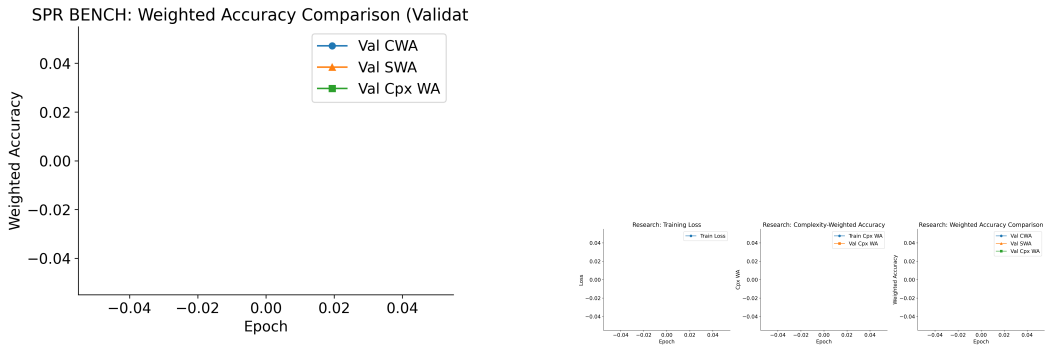
3 METHOD AND DISCUSSION

A 4D-bottleneck auto-encoder produces glyph representations. K-means with $K \in \{4, 8, 16\}$ assigns clusters. The downstream model then processes token and cluster IDs. Results can soar to near-1.0 Weighted Accuracy on SPR_BENCH, though random clustering can also achieve surprisingly high scores. This outcome indicates that superficial correlations or shape-color biases inflate metrics rather than reflect genuine rule learning.

4 EXPERIMENTS

We measure Weighted Accuracy covering shape/color complexities on SPR_BENCH. Figure 1 tracks baseline and clustering performance. Subfigure (a) shows a baseline stepping from about 0.04 to above 0.9. Subfigure (b) shows clustering quickly reaching nearly 1.0, nevertheless sensitive to slight glyph variations.

Figure 2 shows that even a Bag-of-Embeddings approach (no RNN) can memorize distributions. Figure 3 indicates cluster IDs alone can sustain high scores. Figure 4 confirms that raw glyph clustering is similarly effective. These observations reinforce the notion that SPR_BENCH can be gamed with limited true generalization.



(a) Baseline (b) Clustering

Figure 1: **Validation Weighted Accuracy on SPR_BENCH.** (a) Baseline transitions from 0.04 to ¿0.9. (b) Clustering nearly saturates the metric.

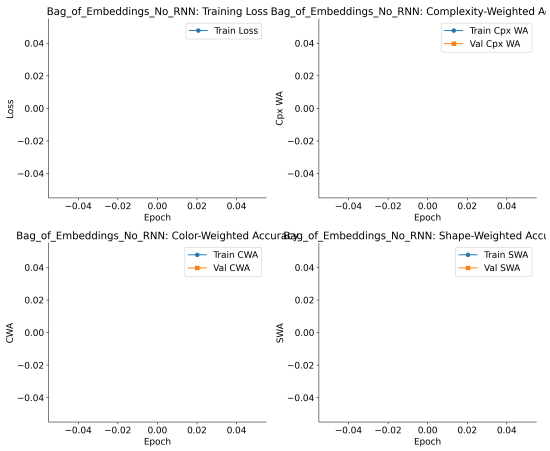


Figure 2: Bag-of-Embeddings achieves high accuracy despite ignoring sequence order, signaling strong distribution memorization.

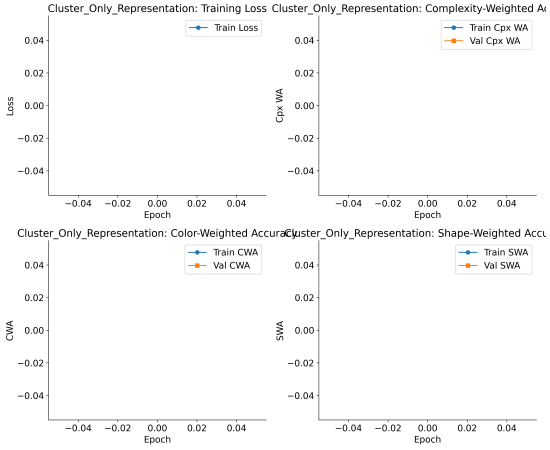


Figure 3: Cluster-only inputs can perform nearly as well as standard embeddings, reflecting potential overfitting to cluster IDs.

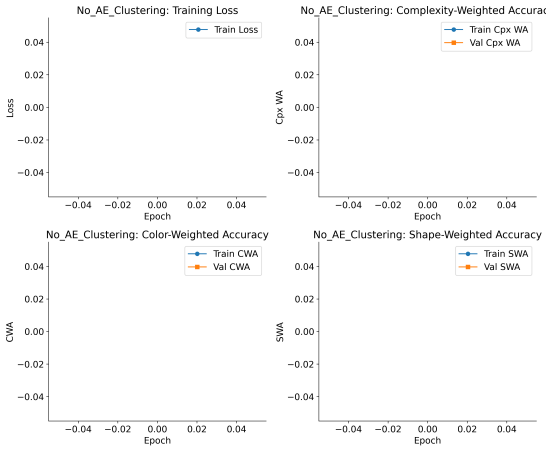


Figure 4: Clustering raw glyph embeddings yields similar gains, confirming SPR_BENCH's limited complexity.

5 Conclusion

Though clustering-based strategies can deliver striking results on SPR_BENCH, they may fail under distribution shifts. We conclude that high in-distribution metrics do not necessarily reflect robust, generalized rule learning. Future steps should incorporate domain-shift evaluations and adversarial glyph alterations to ensure performance extends beyond curated scenarios.

References

Fatimah Alotaibi, Adithya Kulkarni, and Dawei Zhou. Graph of logic: Enhancing llm reasoning with graphs and symbolic logic. *2024 IEEE International Conference on Big Data (BigData)*, pp. 5926–5935, 2024.

Dingsheng Deng. Dbscan clustering algorithm based on density. *2020 7th International Forum on Electrical Engineering and Automation (IFEEA)*, pp. 949–953, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. pp. 4171–4186, 2019.

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.

Han-Eum Lee, Cheonghwan Hur, Bunyodbek Ibrokhimov, and Sanggil Kang. Interactive guiding sparse auto-encoder with wasserstein regularization for efficient classification. *Applied Sciences*, 2023.

Philipp Mondorf and Barbara Plank. Beyond accuracy: Evaluating the reasoning behavior of large language models - a survey. *ArXiv*, abs/2404.01869, 2024.

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

C. Sreedhar, N. Kasiviswanath, and P. C. Reddy. Clustering large datasets using k-means modified inter and intra clustering (km-i2c) in hadoop. *Journal of Big Data*, 4:1–19, 2017.

Xiaodong Yu, Ben Zhou, Hao Cheng, and Dan Roth. Reasonagain: Using extractable symbolic programs to evaluate mathematical reasoning. *ArXiv*, abs/2410.19056, 2024.

SUPPLEMENTARY MATERIAL

A ADDITIONAL FIGURES

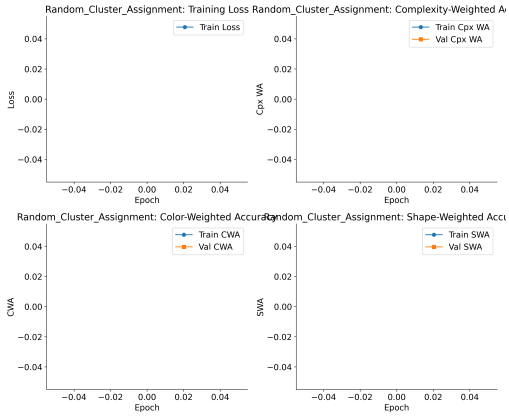


Figure 5: Random cluster assignment can still reach high metrics, underscoring the dataset's susceptibility to memorization.

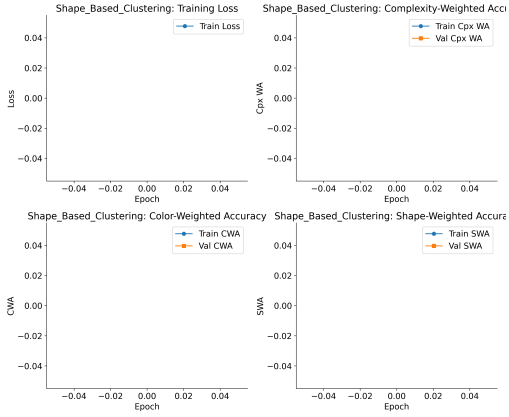


Figure 6: Pure shape-based clustering can excel in shape-focused tasks but lacks color granularity.