CONTEXT-AWARE CONTRASTIVE LEARNING FOR ENHANCED SYMBOLIC PATTERN RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose leveraging context-aware self-supervised contrastive learning to enhance feature representations for the Synthetic PolyRule Reasoning (SPR) task, where symbolic sequences governed by latent logical rules must be classified. Our hypothesis is that integrating advanced data augmentation and denoising strategies can produce representations more robust to positional distortions and noise, thus improving performance on tasks requiring symbolic pattern recognition. We discuss a design for a context-aware contrastive learning framework tailored to symbolic sequences, followed by preliminary experiments on the SPR_BENCH dataset. While the learned embeddings demonstrate promise, results reveal challenges in ensuring that gains generalize across different metrics of shape or color complexity. We discuss pitfalls encountered, such as difficulty in creating effective negative samples for symbolic sequences and the risk of overfitting to fine-grained symbolic traits, to encourage open discussion on ways to refine contrastive methods for real-world symbolic reasoning.

1 INTRODUCTION

Symbolic pattern recognition captures discrete structures often found in logical or rule-based tasks (Goodfellow et al., 2016; Sun et al., 2025). Although deep learning excels in continuous domains, bridging neural networks with symbolic representations remains challenging. Recently, contrastive learning has become a popular technique for learning general-purpose embeddings by bringing semantically similar samples closer together while pushing dissimilar samples apart (Chakraborty et al., 2020; Kim et al., 2024; Choi et al., 2025). However, most prior work focuses on continuous data such as images or time series. Symbolic tasks, particularly those like Synthetic PolyRule Reasoning (SPR), introduce unique hurdles. Sequences are governed by hidden logical rules, and performance can be sensitive to small changes in shape or color tokens. The research here explores whether advanced data augmentation and denoising strategies, specifically adapted for symbolic sequences, can yield broader improvements while highlighting pitfalls encountered in real-world scenarios. By integrating these insights, we hope to encourage systematic studies of context-aware embeddings for symbolic reasoning tasks.

2 RELATED WORK

Contrastive learning has achieved notable success in computer vision, speech, and time-series tasks, typically relying on transformations that preserve semantic features while altering superficial aspects (Chakraborty et al., 2020; Kim et al., 2024; Choi et al., 2025). Symbolic reasoning research has often focused on supervised methods that rely on extensive labeled data (Sun et al., 2025), but these approaches may underperform when labeled examples are scarce or when symbolic rules are too diverse. Our work draws inspiration from these prior approaches but targets symbolic patterns, showing that naive augmentations do not always translate well from continuous to discrete domains. We aim to elucidate challenges in generating context-relevant positive and negative pairs of sequences that vary in shape and color tokens.

3 Background and Method

The Synthetic PolyRule Reasoning (SPR) task comprises symbolic sequences formed from tokens describing shape and color, where hidden logical constraints govern class labels. Performance is evaluated using Shape-Weighted Accuracy (SWA) and Color-Weighted Accuracy (CWA), rewarding models for accurate predictions on sequences that exhibit higher shape or color diversity. The SPR_BENCH dataset provides separate splits for training, development, and testing, while utility metrics highlight intricacies of rule-based classification.

We design a context-aware contrastive learning framework, adapted for discrete symbolic tokens. The core idea is to generate positive and negative pairs based on semantic similarity in shape and color variety. Positive examples undergo controlled transformations such as token masking, token shuffling, or noise injection that preserve overall rule structure, while negative examples differ in critical symbolic features. We pre-train on unlabeled sequences, maximizing agreement between augmented views of the same sequence while minimizing alignment with genuinely dissimilar sequences. A small labeled portion is then used to fine-tune for classification. This approach aims to create embeddings attentive to both local token-level features and global logical structure.

4 Experiments

We evaluate on SPR_BENCH, following a split of 20k training, 5k dev, and 10k test examples. Embeddings are visualized to monitor clustering according to shape and color variety. Despite initial improvements in certain metrics compared to a baseline supervised approach, we observe inconsistent gains on sequences exhibiting high shape diversity. This inconsistency appears partly driven by the difficulty of constructing negative pairs for sequences with overlapping symbolic patterns. Additionally, we note unexpected sensitivity to hyperparameter choices in data augmentation.

While our method approaches or slightly exceeds reported baseline metrics (65.0% SWA and 70.0% CWA), the margin of improvement remains small, suggesting that further refinement is required. Challenges include balancing the complexity of symbolic augmentations with robust generalization and ensuring consistent performance across shape- and color-diverse sequences.

5 Conclusion

We explored a context-aware contrastive learning framework for the SPR task, hypothesizing that symbolic-focused augmentations and denoising could improve representation quality. Empirical findings partially support this hypothesis but also highlight key pitfalls, such as difficulty in constructing negative pairs and sensitivity to symbolic variety. Future directions include automatically learning which symbolic tokens to mask or shuffle, more sophisticated methods for negative sampling, and systematic evaluations across diverse rule-based settings. We hope these observations encourage the community to examine real-world pitfalls in adaptively designing self-supervised methods for symbolic reasoning tasks.

References

Souradip Chakraborty, A. R. Gosthipaty, and Sayak Paul. G-simclr: Self-supervised contrastive learning with guided projection via pseudo labelling. *2020 International Conference on Data Mining Workshops (ICDMW)*, pp. 912–916, 2020.

Jinkyeong Choi, Yejin Noh, and Donghyeon Park. Similarity-guided diffusion for contrastive sequential recommendation. 2025.

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

Dokyun Kim, Sukhyun Cho, Heewoong Chae, Jonghun Park, and Jaeseok Huh. Semi-supervised contrastive learning with decomposition-based data augmentation for time series classification. *Intelligent Data Analysis*, 29:94 – 115, 2024.

Zhong-Hua Sun, Rui Zhang, Zonglei Zhen, Da-Hui Wang, Yong-Jie Li, Xiaohong Wan, and Hongzhi You. Systematic abductive reasoning via diverse relation representations in vector-symbolic architecture. *ArXiv*, abs/2501.11896, 2025.

SUPPLEMENTARY MATERIAL

Additional details, such as hyperparameters, evaluation scripts, and expanded visualizations, are provided here for completeness. For hyperparameter tuning, we performed a random search over: batch size from 64 to 512, learning rate from 1e-4 to 1e-2, and temperature from 0.05 to 0.2. A batch size of 256, a learning rate of 1e-3, and a temperature of 0.1 generally yielded the best trade-off between stability and performance.

Below is a snippet of the dataset loader used in our analyses:

```
def load_spr_bench(root: pathlib.Path) -> DatasetDict:
    # Implementation details loading symbolic sequences
    # ...
    return ds
```

We also explored ablation studies, examining various token-level corruption strategies (e.g., partial shape masking) and more rigorous negative sampling. Due to page constraints, these results are summarized here. Figures illustrating the effects of varying noise levels on symbolic sequences were likewise deferred to this appendix to avoid duplication in the main text.