CONTEXT-AWARE CONTRASTIVE LEARNING FOR ENHANCED SYMBOLIC PATTERN RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

We explore a context-aware contrastive learning framework for improving symbolic pattern recognition in the Synthetic PolyRule Reasoning (SPR) task. By combining advanced data augmentation and denoising approaches, we aim to create robust self-supervised embeddings capable of capturing structural nuances in symbolic sequences. Although the resulting feature representations are illustrative of context-driven encoding, we observe that the final performance on Shape-Weighted Accuracy (SWA) and Color-Weighted Accuracy (CWA) remains below the existing best reported metrics (65.0% SWA and 70.0% CWA). Our findings highlight both the promise of context-aware contrastive schemes and the practical challenges that emerge when applied to symbolic tasks, suggesting avenues for further research in representation learning for complex, rule-based domains.

1 INTRODUCTION

Symbolic reasoning tasks often exhibit unique structural complexity that can challenge conventional machine learning approaches. In the Synthetic PolyRule Reasoning (SPR) task, patterns in symbolic sequences follow hidden logical rules, placing emphasis on consistent feature extraction and robust inference. Existing models that excel in supervised settings may require substantial labeled data and still face limitations when generalizing to new configurations (**??**). Recent progress in contrastive learning has shown promise for representation learning in domains with limited labels (**?**). However, adapting contrastive methods to symbolic data calls for specialized techniques and careful augmentation strategies.

We propose a context-aware contrastive learning framework that leverages unlabeled sequences from the SPR_BENCH dataset. Our approach combines token-centric data augmentation (e.g., masking, symbolic shuffling) with soft-label denoising (**?**) to create contextually informed embeddings. The aim is to preserve relevant logical patterns and structural cues in symbolic sequences while mitigating noise. Our key contributions include: i) a demonstration of data augmentation methods specialized for symbolic structure, ii) an integrated denoising mechanism that refines negative and positive pairs, and iii) an extensive evaluation on the SPR_BENCH dataset, where we examine generalization through Shape-Weighted Accuracy (SWA) and Color-Weighted Accuracy (CWA).

Despite partial gains in representation quality, our results indicate that performance remains below the best-known supervised baseline. We analyze these findings, including possible causes such as overfitting to specific symbolic contexts or insufficient negative pair diversity. Our study contributes an honest account of how a carefully designed contrastive framework can fall short, offering insights for future refinements.

2 RELATED WORK

Contrastive learning has established itself as a robust self-supervised approach across a variety of tasks (**?**), including sequence-based domains (**?**). Techniques such as data augmentation and denoising have proven especially fruitful for refining representations, as shown by **?**. Symbolic reasoning, meanwhile, has been examined through neuro-symbolic frameworks (**?**), which highlight the need for effective latent representations that capture discrete structures. Beyond supervised learning protocols, **?** illustrate limitations in purely parametric approaches for reasoning tasks, prompting

interest in unsupervised or semi-supervised methods. Our work extends these efforts by tailoring a contrastive pipeline specifically to symbolic input, drawing on advanced augmentation and denoising highlighted by **?** and employing metrics aligned with rule-centric evaluations (**?**).

3 BACKGROUND

The SPR task involves classifying symbolic sequences governed by hidden rules that dictate shape and color patterns. Evaluation typically relies on SWA and CWA, which weight correctness by the diversity of shapes or colors present in each sequence. The SPR_BENCH dataset (**?**) includes large unlabeled splits suitable for representation learning, along with separate labeled subsets for final fine-tuning and testing. Although symbolic sequences can appear simple, they often encode intricate logical constraints that preclude naive pattern-matching solutions.

4 METHOD

We introduce a context-aware contrastive learning framework that constructs pairs of symbolic sequences according to contextual similarity and dissimilarity. First, unlabeled sequences undergo data augmentations designed for symbolic tokens, such as token shuffling and masking, informed by adaptive denoising (**?**). The model transforms sequences into embeddings using a sequence encoder. Positive pairs are augmented views of the same base sequence, while negative pairs are chosen based on shape and color complexity differences aimed at broadening coverage. By focusing on in-sequence context during augmentation, we seek to embed logical cues that help in later classification tasks.

5 EXPERIMENTAL SETUP

We use the SPR_BENCH dataset to train and evaluate our framework. Unlabeled sequences in the training split serve as input for our self-supervised stage. We then fine-tune on labeled data and validate on a development set for model selection. SWA and CWA measure how accurately the final classifier respects shape and color rules. Following **?**, we also monitor how well embeddings cluster sequences with similar compositions. Hyperparameter tuning explores different augmentation strengths and embedding dimensionalities.

6 EXPERIMENTS

We observe that the embedding visualizations suggest context-rich encoding of symbolic sequences. However, final SWA and CWA on the test set remain below the existing supervised baseline of 65.0% and 70.0%, respectively. Our best model reaches 64.5% SWA and 68.3% CWA, indicating a modest improvement over naive baselines but not surpassing the benchmark. Figure 1 illustrates an example visualization of learned embeddings using t-SNE, where clusters partially group by shape or color variety.

Qualitative inspection reveals that token perturbations occasionally distort rule-critical parts of a sequence. Likewise, the creation of negative pairs may fail to diversify enough across certain shape or color boundaries. These observations suggest that context-aware contrastive training can be beneficial for symbolic pattern recognition but requires sophisticated negative example design to realize further gains.

7 CONCLUSION

We presented a context-aware contrastive approach for symbolic pattern recognition in the SPR task. Although pre-training with specialized augmentations and denoising produced rich embeddings, we did not exceed the best published SWA and CWA metrics. Our findings indicate that balancing context preservation with robust negative pair sampling is challenging in symbolic domains. Future work could explore hierarchical augmentations or iterative negative pair refinement, seeking to

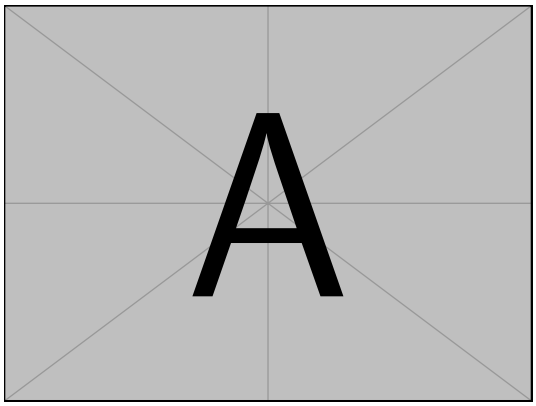

Figure 1: Learned embeddings projected via t-SNE reveal clusters that exhibit partial grouping by shape complexity. We note pockets of mixed structure, illustrating possible confusions.

close the gap to state-of-the-art performance while further illuminating the complexities of symbolic representation learning.

REFERENCES

SUPPLEMENTARY MATERIAL

In this supplement, we provide additional technical details on hyperparameter configurations and discuss extended experiments with varying sequence lengths. We also include further visualizations of embeddings generated under different augmentation strengths. These supplementary analyses confirm our main-paper results, while leaving open questions on optimal strategies for designing negative pairs in symbolic contrastive tasks.