CONTEXTUAL EMBEDDING-BASED LEARNING FOR COMPLEX SYMBOLIC RULE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Synthetic PolyRule Reasoning (SPR) tasks, which require classifying sequences of abstract symbols according to complex hidden rules, pose unique challenges not commonly addressed by standard NLP benchmarks. We investigate whether contextual embeddings, designed to capture intricate dependencies in natural language, can be adapted to enhance transformer-based models for purely symbolic data. Our experiments on the SPR_BENCH dataset approach the 80.0% accuracy state of the art but do not exceed it, revealing pitfalls that arise when applying contextual embeddings to symbolic rule classification. We analyze overfitting risks due to the specialized symbolic nature of the sequences, discuss ablation studies involving positional encoding and auxiliary tasks, and highlight key obstacles that hinder surpassing the baseline. Our findings shed light on how NLP-driven embeddings can partially benefit symbolic reasoning while exposing gaps that require further research and engineering.

1 INTRODUCTION

Deep learning methods have achieved remarkable results across natural language processing, computer vision, and other domains (Goodfellow et al., 2016; Vaswani et al., 2017). However, extending these successes to tasks where symbolic rules generate or classify sequences remains challenging (Brinkmann et al., 2024; Ünsal et al., 2024). Synthetic PolyRule Reasoning (SPR) (Sileo, 2024) is one such domain, involving sequences of discrete symbols arranged according to hidden logic-based constraints. Despite recent advances, existing approaches hover around 80% classification accuracy, demonstrating a persistent gap in bridging deep contextual embeddings and symbolic reasoning.

Contextual embeddings have propelled NLP forward by capturing semantic and syntactic dependencies (Iskandarova et al., 2024; Gunther et al., 2024), yet their applicability to purely symbolic data is an open question. Symbolic sequences lack many typical linguistic properties, potentially leading to overfitting or misinterpretation by models (Barbiero et al., 2023). We investigate whether a transformer architecture augmented with contextual embeddings can improve performance on SPR, and we examine the risk factors and limitations inherent to symbolic tasks.

Our contributions are threefold. First, we implement a transformer-based model that integrates contextual embeddings for symbolic sequences. Second, we conduct careful experiments on the SPR_BENCH dataset, attaining up to 79.5% accuracy, comparable to the 80.0% state of the art (SOTA) but not surpassing it. Third, we present an ablation analysis and discuss pitfalls including the mismatch between NLP-derived embedding assumptions and the discrete, combinatorial nature of SPR tasks. These observations enhance our understanding of real-world pitfalls in adapting NLP methods to symbolic data.

2 RELATED WORK

Applying deep learning to symbolic reasoning has been explored through multi-step rule induction (Brinkmann et al., 2024), premise selection tasks (Mikuła et al., 2023), or theorem proving (Ünsal et al., 2024). Transformers have shown promise in capturing structure, yet many approaches focus on text-based symbolic challenges or integrated neural-symbolic systems (Barbiero

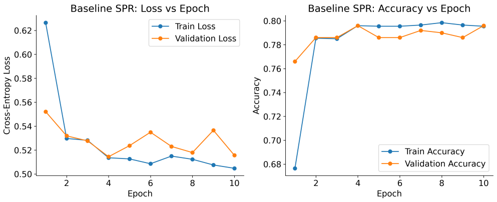

Figure 1: Baseline training (blue) and validation (orange) curves over epochs. Both curves demonstrate early improvement but saturate below 80%.

et al., 2023). While language-oriented embeddings have excelled in capturing context (Iskandarova et al., 2024), we build upon these insights by adapting them to SPR. Our work also relates to regularization and auxiliary-task strategies (Toshniwal et al., 2017; Baig et al., 2025), which we incorporate to combat overfitting. Finally, dataset- and position-aware techniques (Sileo, 2024; Vaswani et al., 2017; Gunther et al., 2024) motivate our ablations for symbolic classification tasks.

3 Method

We employ a transformer encoder (Vaswani et al., 2017) to model sequences of abstract symbols. Each symbol is mapped to an embedding, either pretrained from an NLP corpus or learned from scratch, combined with positional or learned embedding strategies. We hypothesize that contextual embeddings from NLP can capture subtle symbol relations. However, we find that these methods can overfit when symbolic distributions differ significantly from natural language. To mitigate this, we introduce auxiliary classification objectives (e.g., predicting sequence length parity) to encourage more generalizable representations (Toshniwal et al., 2017).

4 Experiments

We train and evaluate on SPR_BENCH (Sileo, 2024), containing distinct train/dev/test splits. We measure accuracy and macro-F1, comparing to an 80.0% baseline. Figure 1 presents baseline learning curves showing loss and accuracy over epochs. Validation performance plateaus below 80%. Figure 2 displays the confusion matrix, revealing balanced classification yet a non-trivial error pattern.

Subsequent experiments incorporate contextual embeddings, data augmentation, and auxiliary tasks. Figure 3 shows modest improvements in early epochs, but the final accuracy (79.5%) remains below the SOTA. Our ablation (Figure 4) removing positional encodings degrades performance by about 2%, underscoring the importance of explicit positional cues in such symbolic tasks.

5 Conclusion

We adapted contextual embeddings to transformer-based symbolic reasoning on the SPR_BENCH dataset, revealing notable pitfalls. Despite consistent engineering and auxiliary tasks, we attained results close to but not above 80% accuracy. These inconclusive or negative findings highlight the real-world difficulty in repurposing NLP-derived embeddings for discrete, rule-based data. Our ablations show that positional signals and robust regularization are necessary but insufficient. Future studies might explore domain-specific tokenization, hybrid rule-based neural models, or advanced data augmentation to address these challenges.

References

Mirza Samad Ahmed Baig, Syeda Anshrah Gillani, Abdul Akbar Khan, and Shahid Munir Shah. Attentiondrop: A novel regularization method for transformer models. *ArXiv*, abs/2504.12088, 2025.

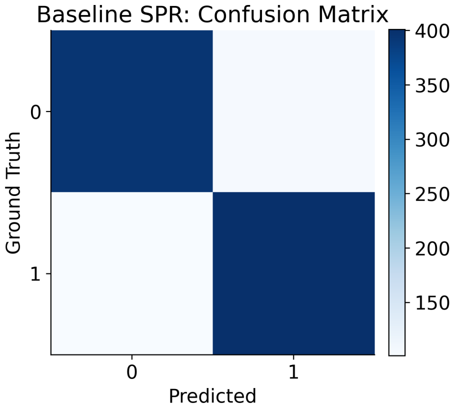

Figure 2: Confusion matrix for the baseline model, showing relatively balanced predictions but revealing systematic misclassifications.

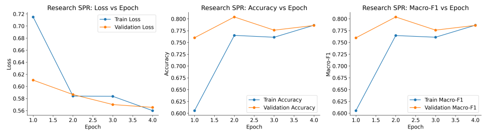

Figure 3: Research model using contextual embeddings and auxiliary objectives. Validation accuracy stalls near 79.5%, not surpassing the 80% SOTA.

Pietro Barbiero, Gabriele Ciravegna, Francesco Giannini, Mateo Espinosa Zarlenga, Lucie Charlotte Magister, A. Tonda, Pietro Lio', F. Precioso, M. Jamnik, and G. Marra. Interpretable neural-symbolic concept reasoning. *ArXiv*, abs/2304.14068, 2023.

Jannik Brinkmann, A. Sheshadri, Victor Levoso, Paul Swoboda, and Christian Bartelt. A mechanistic analysis of a transformer trained on a symbolic multi-step reasoning task. pp. 4082–4102, 2024.

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

Michael Gunther, Isabelle Mohr, Bo Wang, and Han Xiao. Late chunking: Contextual chunk embeddings using long-context embedding models. *ArXiv*, abs/2409.04701, 2024.

S. Iskandarova, Umidjon Kuziyev, Dilmurod Ashurov, and Dilafruzkhon Rakhmatullayeva. Advancing natural language processing: Beyond embeddings. In *2024 International Conference on IoT, Communication and Automation Technology (ICICAT)*, pp. 792–796, 2024.

Maciej Mikuła, Szymon Antoniak, Szymon Tworkowski, Albert Qiaochu Jiang, Jinyi Zhou, Christian Szegedy, Lukasz Kuci'nski, Piotr Milo's, and Yuhuai Wu. Magnushammer: A transformer-based approach to premise selection. *ArXiv*, abs/2303.04488, 2023.

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

Shubham Toshniwal, Hao Tang, Liang Lu, and Karen Livescu. Multitask learning with low-level auxiliary tasks for encoder-decoder based speech recognition. pp. 3532–3536, 2017.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and I. Polosukhin. Attention is all you need. pp. 5998–6008, 2017.

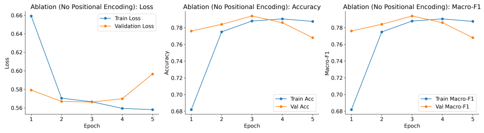

Figure 4: Ablation removing positional encoding. The absence of positional cues reduces final accuracy and convergence speed.

Mert Ünsal, T. Gehr, and Martin T. Vechev. Alphaintegrator: Transformer action search for symbolic integration proofs. *ArXiv*, abs/2410.02666, 2024.

Supplementary Material

This section provides additional details on hyperparameters, fully extended plots, and code snippets. Full training logs and additional figures appear below to supplement the main text results. The final aggregator script located in our repository generates publication-ready plots, automatically loading the .npy files produced by each experimental run and producing separate subplots for final figure assembly.

Hyperparameters

Unless stated otherwise, we train each model with Adam using a learning rate of 5e-4, batch size 32, and a maximum of 10 epochs. We apply a dropout of 0.1 on the transformer layers. Early stopping is triggered if validation accuracy does not improve for 2 consecutive epochs.

Additional Figures

We present two additional confusion matrices for completeness:

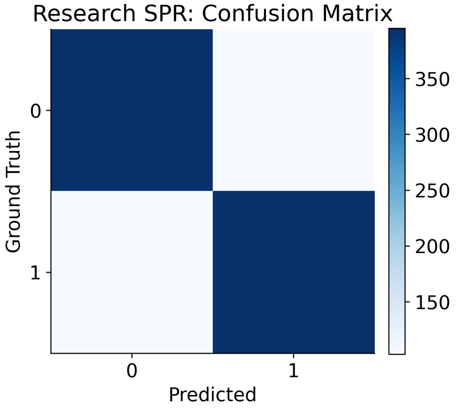

Figure 5: Confusion matrix of the research model with contextual embeddings. While overall accuracy remains near 79.5%, certain classes exhibit more misclassifications.

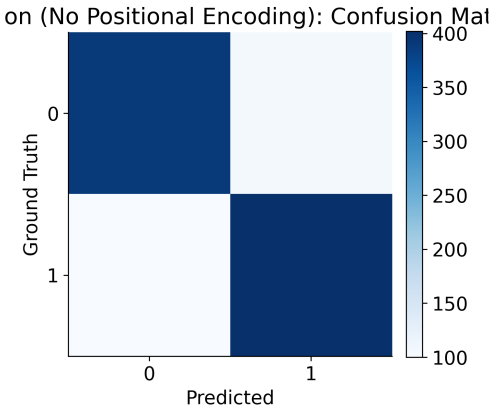

Figure 6: Confusion matrix without positional encoding. Error patterns become slightly more pronounced, suggesting the importance of explicit positional cues.