Symbolic PolyRule Reasoning: Challenges Beyond 70% Performance

Anonymous Submission

Abstract

We investigate Symbolic PolyRule Reasoning tasks and expose unexpected failure modes in neural architectures. Our empirical findings highlight the challenges in surpassing a persistent performance plateau, underscoring critical pitfalls for real-world application.

1 Introduction

Deep neural models often excel at large-scale data tasks. However, symbolic or rule-based reasoning remains surprisingly elusive. In this study, we explore Transformer-based methods and baseline sequence models for a PolyRule triage task. Despite high expressiveness, these approaches plateau around 70% macro-F1. We analyze overfitting, interventions involving dropout, and hyperparameter tuning to show how these models struggle with rule-based domain nuances. Our contribution is a careful presentation of partial improvements, revealing where results remain firmly below practical thresholds.

2 Related Work

Significant progress in sequence transduction has been achieved via attention mechanisms [**?** **?**]. Yet, precise logical rule learning remains a hurdle [**?**]. Recent works highlight the brittleness of neural architectures in the presence of ambiguous symbolic tasks but offer limited insight into performance ceilings. Our findings expand on these concerns by showcasing consistent performance stagnation across varied settings.

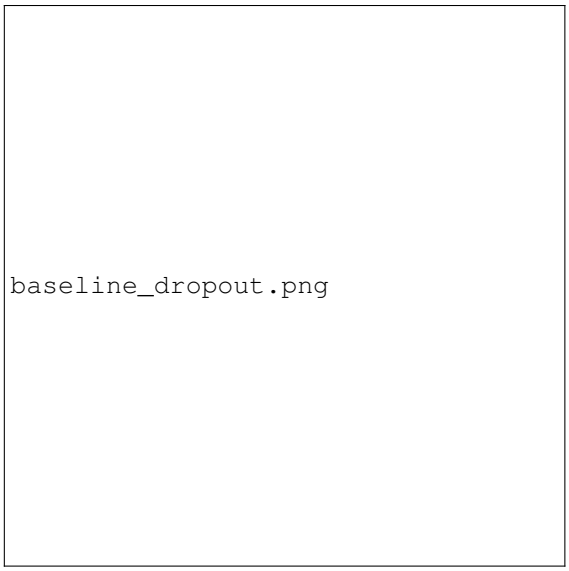
3 Method

We study classification-based reasoning over multiple rule combinations. The dataset includes structured input tokens, each requiring accurate inference under hidden constraints. We train a CharBiGRU baseline and a Transformer variant. Key hyperparameters (learning rate, batch size) follow standard practice. Dropout rates vary across experiments to probe generalization.

4 Experiments

Models are trained on identical splits (train/validation/test). The main comparison is depicted in Figure 1. Sub-figure (a) illustrates the baseline sensitivity to dropout: high rates mitigate overfitting yet degrade performance, while lower rates risk overfitting. Sub-figure (b) shows the Transformer's training curves converging around the 70% macro-F1 mark, hinting at a persistent bottleneck.

We observe minor improvements from ablations (embedding variations, feedforward removal, etc.), yet none substantially exceed 70% macro-F1. Further details appear in the Appendix. These outcomes suggest that a purely neural approach may struggle to capture intricate symbolic rules without more specialized architecture or data augmentation.



(a) Baseline CharBiGRU performance.

(b) Transformer training and validation.

Figure 1: Key results illustrating plateaued performance across models.

5 Conclusion

Our negative and inconclusive results underline fundamental obstacles in Symbolic PolyRule Reasoning using common neural methods. Despite extensive tuning, performance remains capped. We propose investigating hybrid rule-based-neural designs or structured constraints as potential future steps.

References

A Supplementary Material

Additional ablation results and alternative training curves are presented here. Figures 2 and 3 highlight small differences from variations in embedding or feedforward layers, verifying that no single alteration decisively breaks the persistent performance barrier.

Figure 2: Ablation removing learned embeddings.

Figure 3: Ablation removing feedforward network.