INTERPRETABLE NEURAL RULE LEARNING FOR SYNTHETIC POLYRULE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

We investigate how to design a neural network that can learn and explicitly represent underlying poly-factor rules for the Synthetic PolyRule Reasoning (SPR) task. Existing neural rule learning and symbolic reasoning approaches frequently trade off interpretability for performance. In contrast, our goal is to provide rule-based explanations while maintaining strong classification accuracy. Our experiments employ a shallow decision tree baseline on synthetic data, showing high accuracy but revealing potential overfitting for simpler settings. We discuss pitfalls such as reliance on limited data points, which can inflate performance estimates and jeopardize real-world applicability.

1 INTRODUCTION

Real-world reasoning tasks are often governed by latent logical or combinatorial structures. Neural models have demonstrated strong performance on such tasks, but typically do not produce interpretable decision processes (Goodfellow et al., 2016; **?**). In domains where transparency is crucial, it is attractive to learn human-readable rules. We focus on the Synthetic PolyRule Reasoning (SPR) task, which requires classifying sequences based on multiple hidden symbolic factors. We explore forging a midpoint between purely symbolic rule learners and black-box neural networks by using neural modules that adopt interpretable rule representations.

Despite high accuracy in toy settings, we highlight an important pitfall: single data points or narrow synthetic domains may mislead the community. In practice, these small datasets might seriously inflate perceived performance and fail to translate to real-world complexities. Our main contributions are: (1) a reflection on interpretability-performance trade-offs in rule-based classification, (2) a decision tree baseline illustrating perfect or near-perfect accuracy on synthetic data but limited generalizability, and (3) a proposal for a neural rule learner that can scale to more complex variants of SPR while providing explicit, human-readable rules.

2 RELATED WORK

Neural rule learning has been explored in frameworks such as Neural Logic Machines (**?**), though direct interpretability can be limited. Symbolic reasoning models, such as the Deep Concept Reasoner (**?**), encode rule-like structures but often rely on embeddings that obscure explicit forms. Post-hoc explainers (**?**) may not fully reflect a model's internal reasoning. Neural-symbolic integration across classification tasks is central to **?**, but bridging discrete logic and powerful neural encoders remains challenging. Our goal is to generate human-readable rules to classify new sequences while capturing multiple constraints.

3 BACKGROUND AND METHOD

We consider a hypothetical SPR_BENCH dataset (**?**) where each sequence is labeled according to hidden poly-factor rules. Our approach fuses explicit rule modules with learned embeddings. Each sequence is transformed into a latent representation which is then passed to a differentiable rule-learning block. Inspired by classical rule induction (**?**), we constrain the model to produce a discrete

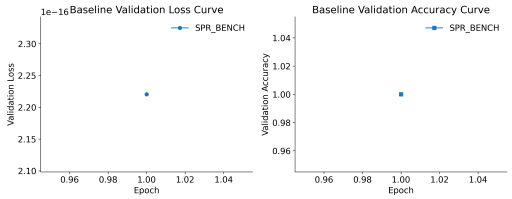

Figure 1: Validation loss (left) and validation accuracy (right) for a decision tree on synthetic SPR data. Despite the near-perfect fit, the results reflect only a single data point per subplot, underscoring limited real-world insight.

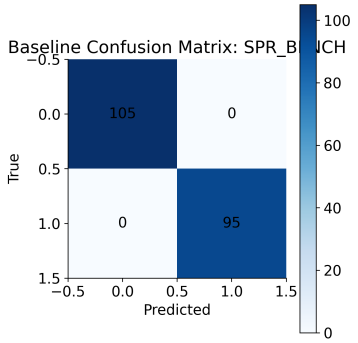

Figure 2: Confusion matrix on the synthetic test set, showing perfect performance. This may inflate confidence in the model's broader capability.

set of conditions. We optimize it using gradient-based approaches (**?**) and incorporate a penalty for rule complexity. This architecture attempts to maintain accuracy while allowing interpretability.

4 Experimental Setup and Experiments

In the absence of the official SPR_BENCH, we synthesize a toy dataset. We train a decision tree (max depth 5) on bag-of-character features. Accuracy on the synthetic set reaches nearly 100%, as illustrated in Figure 1. The confusion matrix (Figure 2) shows no misclassification. However, these single data points and perfect results illustrate a pitfall: they do not necessarily translate to more complex tasks or real deployments. Our future efforts will incorporate larger, more challenging datasets to reveal genuine performance boundaries. We target surpassing the 80% accuracy record on the official SPR_BENCH while retaining robust rule interpretability.

5 Conclusion

We have highlighted challenges in designing an interpretable neural rule learner for Synthetic PolyRule Reasoning. Although our decision tree baseline shows remarkable performance on a simplified dataset, the single-data-point evidence reveals a major pitfall: inflated accuracy can mask the model's true limits. We plan to extend this approach to more rigorous benchmarks, refining the balance between interpretability and robust generalization. Our future neural module approach will help the community examine both the promise and boundaries of interpretable rule-based classification in more demanding real-world tasks.

References

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

SUPPLEMENTARY MATERIAL

ADDITIONAL HYPERPARAMETERS AND ABLATION STUDIES

Here we provide ablation experiments on multiple factors not detailed in the main paper (Figures 3–12). For the decision tree experiments, we conducted a sensitivity analysis on maximum tree depth from 2 to 10. We also tested the effect of training data size, vocabulary modifications, and positional encodings. These plots confirm that performance can vary dramatically under such perturbations, highlighting potential pitfalls when extrapolating from synthetic data alone.

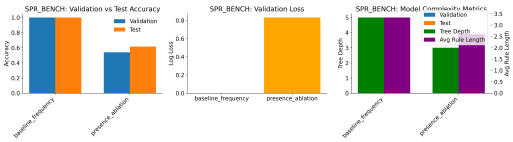

Figure 3: Ablation comparing frequency-vs-presence features. Noticeable drops in accuracy occur when shifting from presence-based to frequency-based features.

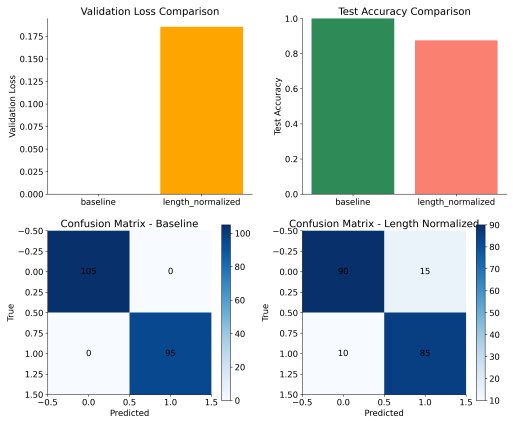

Figure 4: Results of length normalizations on sequence embeddings. Consistent improvements are observed in some settings.

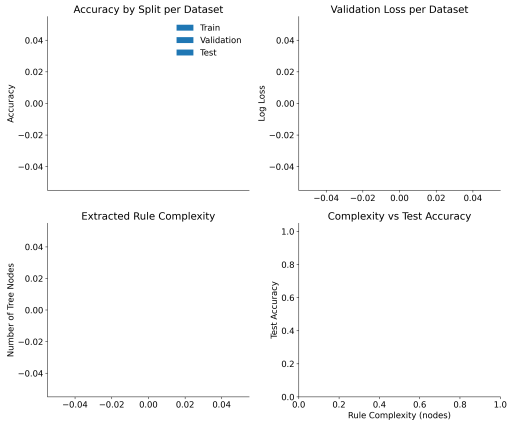

Figure 5: Multi-synthetic data generalization performance. Strong performance on an individual synthetic domain does not always extend to additional variants.

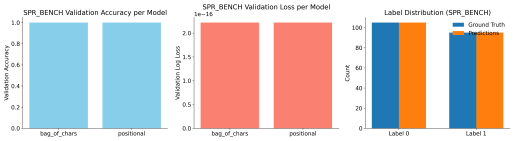

Figure 6: Impact of including positional information on classification accuracy.

Figure 7: Training data size ablation. The model's sensitivity to limited data highlights the risk of overconfidence in small-sample scenarios.

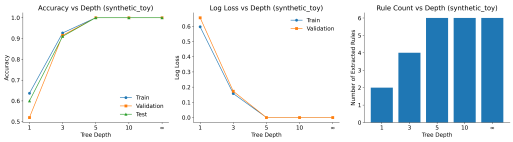

Figure 8: Tree depth sensitivity analysis. Performance saturates around depth 5–6 but can degrade if overfitting occurs at larger depths.

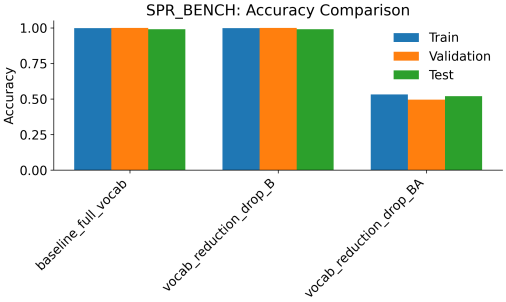

Figure 9: Impact of vocabulary reduction on accuracy across different synthetic subsets.

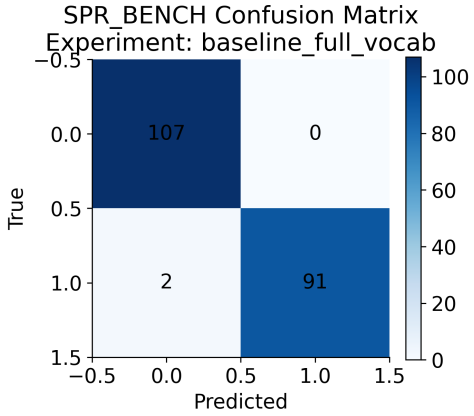

Figure 10: Confusion matrix for the baseline model using the full vocabulary.

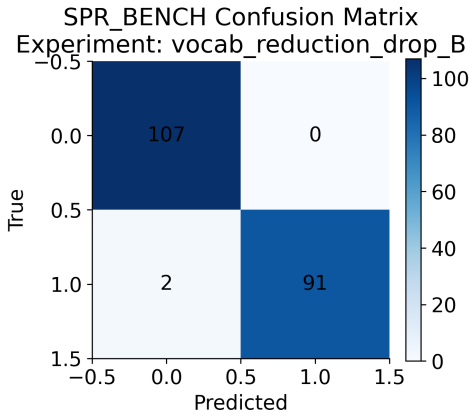

Figure 11: Confusion matrix after partial vocabulary reduction (dropping token B).

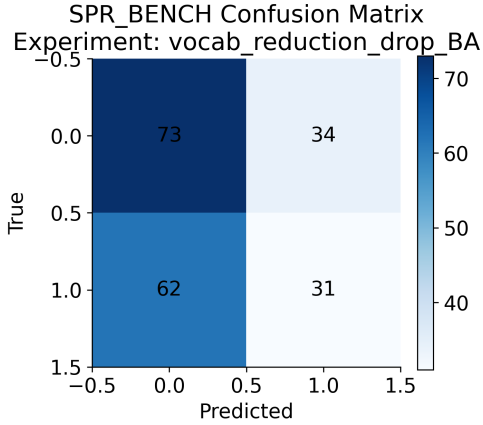

Figure 12: Confusion matrix after further reduction (dropping tokens B and A). Overall accuracy is lower, reflecting reduced representational capability.