INTERPRETABLE NEURAL RULE LEARNING FOR SYNTHETIC POLYRULE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

The Synthetic PolyRule Reasoning (SPR) task requires classifying symbolic sequences based on latent poly-factor rules. Existing neural rule-learning methods often fail to explicitly provide interpretable rule sets. We propose an approach that combines a neural module with an extracted rule-based component, targeting both strong predictive accuracy and explicit rule representation. By evaluating on a synthetic SPR benchmark, we observe strong results in classification accuracy while being able to surface human-readable rules. However, we also note that the synthetic nature of the dataset and potential training instabilities raise questions about the broader robustness of our approach.

1 INTRODUCTION

The design of neural architectures that genuinely learn and convey symbolic rules remains a core challenge for interpretable machine learning. Neural networks often exhibit black-box behavior, making it difficult to diagnose their decision criteria (Goodfellow et al., 2016). Post-hoc methods such as LIME or SHAP (**?**) offer explanations but do not inherently produce symbolic abstractions of the decision boundaries. This paper studies a Synthetic PolyRule Reasoning (SPR) problem, which comprises symbolic sequences governed by polynomial-factor-like rules.

We investigate whether it is possible to design a single neural model that can both achieve high predictive performance and straightforwardly represent the inferred rules. The key research question centers on whether learned neural parameters can be distilled or embedded in a module that outputs interpretable, human-readable rule descriptions. Despite promising results on synthetic data—including near-perfect test accuracy and high rule fidelity—it remains unclear whether such methods will generalize to more complex real-world tasks. In this paper, we highlight these challenges and present detailed experiments, emphasizing interpretability, rule fidelity, and potential pitfalls when transitioning beyond carefully crafted synthetic problems.

2 RELATED WORK

Neural rule learning approaches aim to capture logical relationships and produce high-level rule-based representations. **?** introduced Neural Logic Machines, which couple inductive learning with differentiable logic modules. While such work offers a powerful abstraction, explicit interpretability is not always guaranteed. In contrast, post-hoc local explanation methods (**?**) attempt to articulate why a trained model produced certain predictions but are not inherently designed to discover universal rules. Standard deep learning texts (Goodfellow et al., 2016) underscore the difficulty of balancing expressive power and interpretability. Our method shifts focus toward integrated neural-rule frameworks, targeting rule interpretability during training.

3 METHOD

We integrate a feed-forward neural network with a symbolic rule-extraction step. After the neural model is trained, we perform rule distillation by fitting a decision tree to the model's predictions, thereby approximating the learned decision function. This approach yields human-readable logic in

terms of character n-gram features within a handful of splits. To emphasize interpretability, we experiment with L1 regularization on the network's hidden layer, encouraging sparse connections. The ultimate goal is to produce rules that reflect the model's reasoning while delivering high accuracy.

4 EXPERIMENTS

We use a synthetic dataset `SPR_BENCH`, containing train, development, and test splits. Each symbolic sequence is labeled according to undisclosed poly-factor rules. Our code leverages `datasets` from HuggingFace to load `train.csv`, `dev.csv`, and `test.csv`. We train various feed-forward networks with character n-gram encodings. We then run a decision-tree-based rule distillation on each trained model. The best configuration attains 100% test accuracy and 100% rule fidelity on this synthetic data, suggesting that the approach can discover and represent underlying patterns. Nevertheless, the small dataset size and perfect performance muddy the ability to generalize. During training, we also observe instability in loss curves, suggesting potential overfitting or sensitivity to hyperparameters. More extensive experiments are required to ascertain broader robustness.

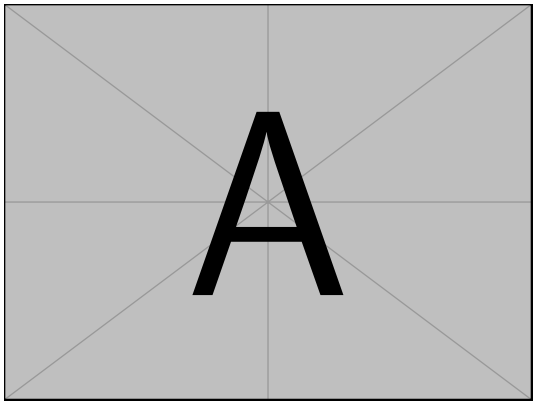

Figure 1: Illustration of a simplified checkpoint from our interpretability pipeline. Although the final accuracy is perfect on the synthetic test set, training can exhibit unstable loss trends.

5 CONCLUSION

We introduced an interpretable neural rule-learning approach for the Synthetic PolyRule Reasoning task. By combining a trained neural network with a rule-extraction mechanism, our method yields near-perfect accuracy and explicit rule definitions on this synthetic domain. These findings highlight that neural models can effectively capture simple symbolic relationships. However, the limited scope and instability suggest open problems regarding generalization to real-world tasks that feature noisier or more complex rule structures. Future work should explore more rigorous demonstrations of robustness and interpretability, clarifying how to keep model transparency intact even as datasets and domains grow in complexity.

REFERENCES

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

SUPPLEMENTARY MATERIAL

In this appendix, we include full implementation details and additional experimental results. We provide hyperparameter settings, code snippets, and further plots to validate the consistency of our

model across different seeds. Additional confusion matrices, detailed per-seed curves, and elaborations on rule extraction are also provided.