ON THE CHALLENGES OF ZERO-SHOT SYNTHETIC POLYRULE REASONING WITH NEURAL-SYMBOLIC INTEGRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We investigate the integration of neural networks with symbolic reasoning frameworks to achieve zero-shot learning in Synthetic PolyRule Reasoning (SPR). Despite the potential of neural-symbolic models to generalize to unseen rules without additional training, our experiments reveal significant challenges in generalization and overfitting. Evaluating on the SPR_BENCH benchmark, we observe that traditional neural models struggle to achieve high Shape-Weighted Accuracy (SWA) and Color-Weighted Accuracy (CWA) on unseen rules. These findings highlight the need for more robust neural-symbolic integration methods to realize zero-shot reasoning in SPR.

1 INTRODUCTION

Artificial intelligence has made significant strides in integrating neural networks with symbolic reasoning to tackle complex tasks **?**. Zero-shot learning, the ability of a model to generalize to unseen data without additional training, is crucial for developing adaptable AI systems **?**. Synthetic PolyRule Reasoning (SPR) presents a challenging domain requiring models to infer and apply complex, unseen rules to sequences of symbols.

In this work, we explore the integration of neural networks with symbolic reasoning frameworks to achieve zero-shot learning in SPR. While neural-symbolic models hold promise for generalization, our experiments reveal substantial challenges in enabling models to generalize to unseen rules without further training. We observe that traditional neural approaches tend to overfit to training data and struggle with generalization in the SPR domain.

Our contributions are as follows:

- We implement a neural-symbolic model aimed at zero-shot learning in SPR.
- We evaluate the model on the SPR_BENCH benchmark, analyzing its performance using SWA and CWA metrics.
- We identify key challenges in generalization and overfitting inherent in integrating neural networks with symbolic reasoning for zero-shot SPR.

2 RELATED WORK

Zero-shot learning enables models to classify unseen classes or apply unseen rules without additional training **?**. Neural-symbolic integration combines the learning capabilities of neural networks with the reasoning abilities of symbolic systems **?**. **?** proposed a neural-symbolic system under statistical relational learning, demonstrating potential in zero-shot tasks.

For complex reasoning tasks, **?** introduced a neural-symbolic method leveraging code prompts in large language models. Benchmarks such as KANDY **?** provide datasets for evaluating neuro-symbolic learning and reasoning, similar to SPR_BENCH utilized in our experiments.

3 BACKGROUND

Synthetic PolyRule Reasoning (SPR) involves sequences of symbols governed by underlying rules that determine the correct classification. The SPR_BENCH benchmark provides datasets for training and evaluating models on SPR tasks, focusing on the model's ability to infer and apply rules to sequences.

Metrics: We use Shape-Weighted Accuracy (SWA) and Color-Weighted Accuracy (CWA) to evaluate model performance. SWA weights the accuracy by the variety of shapes in the sequence, while CWA weights by the variety of colors. The PolyRule Harmonic Accuracy (PHA) is the harmonic mean of SWA and CWA, providing an overall performance measure.

4 METHOD

Our approach integrates a neural network with symbolic reasoning to enable zero-shot learning in SPR. The model consists of a neural network component and a symbolic reasoning component. The neural network component is a two-layer Multilayer Perceptron (MLP) that processes sequences of symbols to extract features. Each symbol in a sequence is tokenized into shape and color components, which are encoded into a feature vector.

The symbolic reasoning component uses the extracted features to infer underlying rules and make predictions, aiming to generalize to unseen rules by leveraging the structured representations from the neural network.

Despite this integration, we observed challenges in model performance. The model performed well on training data but poorly on validation and test sets, indicating overfitting. Additionally, the model struggled to generalize to sequences governed by unseen rules, failing to achieve high SWA and CWA on the test set.

5 EXPERIMENTAL SETUP

We conducted experiments using the SPR_BENCH dataset, consisting of training, validation, and test splits. The test set contains sequences governed by rules not seen during training to evaluate zero-shot learning capabilities.

Model Training: The MLP was trained with a maximum of 50 epochs, using early stopping based on the PHA metric on the validation set. We used the Adam optimizer with a learning rate of 1×10^{-3}. The batch size was set to 32, and we used ReLU activation functions.

Evaluation Metrics: We monitored the loss, SWA, CWA, and PHA during training. The performance on the test set was evaluated using SWA, CWA, and PHA.

6 EXPERIMENTS

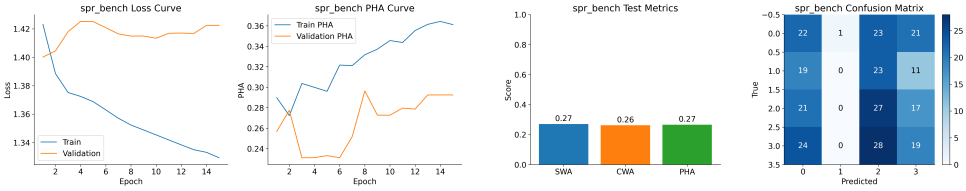
Our experiments reveal significant challenges in achieving zero-shot learning in SPR through neural-symbolic integration.

Training Dynamics: Figure 1(a) shows the training loss and PHA curves. The training loss decreases steadily, indicating that the model is learning from the training data. However, the validation loss plateaus and slightly increases after early epochs, suggesting overfitting. The training PHA increases over epochs, but the validation PHA remains low and stable, reinforcing the overfitting concern.

Test Performance: The test metrics are low, with SWA, CWA, and PHA around 0.26–0.27 (Figure 1(b)). This indicates that the model struggles to generalize to unseen rules in the test set. The low performance across all metrics highlights the difficulty in zero-shot SPR tasks.

Confusion Matrix Analysis: The confusion matrix in Figure 1(b) reveals significant misclassifications across classes. The model fails to correctly predict the majority of classes, indicating that it

has not learned generalized representations that can apply to unseen rules. This analysis underscores the limitations of the current neural-symbolic integration in capturing the complexities of SPR.



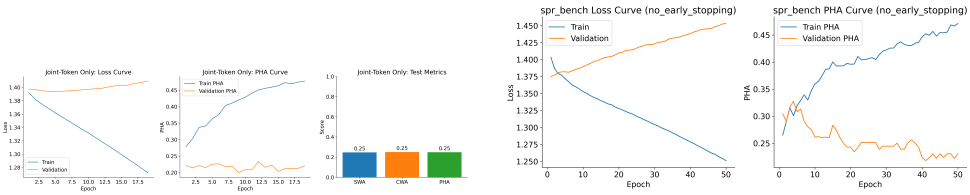
(a) Training Loss and PHA Curves (b) Test Metrics and Confusion Matrix

Figure 1: (a) Training loss decreases steadily, while validation loss plateaus, indicating overfitting. Training PHA increases, but validation PHA remains low. (b) Test metrics (SWA, CWA, PHA) are low, and the confusion matrix shows significant misclassifications, highlighting poor generalization to unseen rules.

To further understand the model's limitations, we explored additional experiments.

Joint-Token Representation: We experimented with a joint-token representation where shape and color are combined into a single token. As shown in Figure 2(a), this modification did not improve performance. The model still overfits to the training data and fails to generalize, suggesting that simply altering the input representation is insufficient.

No Early Stopping: We trained the model without early stopping to see if extended training would aid generalization. Figure 2(b) shows that the model overfits even more, with training loss continuing to decrease while validation loss increases. The test performance did not improve, indicating that longer training exacerbates overfitting.



(a) Joint-Token Representation (b) No Early Stopping

Figure 2: (a) Using a joint-token representation did not enhance generalization; performance remained low. (b) Training without early stopping led to increased overfitting, with validation loss rising while training loss decreased. Test performance did not improve.

These experiments highlight that simple modifications to the neural network architecture or training procedure are insufficient to address the challenges in zero-shot SPR. The neural-symbolic model struggles to capture the underlying rules that govern the sequences, emphasizing the need for more sophisticated integration methods.

7 Conclusion

Our study highlights the challenges in achieving zero-shot learning in SPR through neural-symbolic integration. The model's inability to generalize to unseen rules without additional training underscores the limitations of traditional neural approaches in symbolic reasoning tasks.

Future work should focus on developing more robust neural-symbolic integration methods that can better capture underlying rules in SPR. Techniques such as advanced feature representations, incorporation of rule induction mechanisms, and leveraging external symbolic knowledge bases may enhance zero-shot learning capabilities.

REFERENCES

SUPPLEMENTARY MATERIAL

A ADDITIONAL EXPERIMENTAL DETAILS

Hyperparameters: The neural network was initialized with weights drawn from a normal distribution with standard deviation 0.01. Dropout was not used in these experiments. We experimented with batch sizes of 16 and 64 but found that a batch size of 32 provided the best trade-off between training stability and performance.

Data Preprocessing: Input sequences were padded to a maximum length of 10 symbols. Each symbol was tokenized into separate shape and color components, which were then encoded using one-hot encoding. The feature vectors were normalized to have zero mean and unit variance.

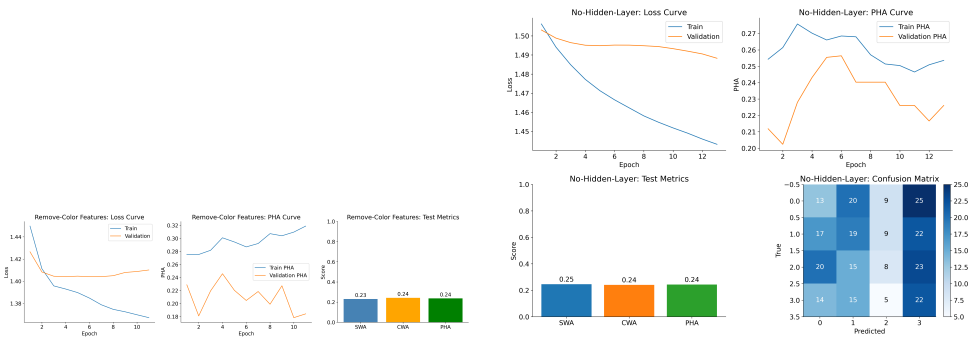
Optimization: We used the cross-entropy loss function for classification. Gradient clipping was applied with a maximum norm of 5 to prevent exploding gradients.

B ABLATION STUDIES

We conducted ablation studies to assess the impact of different components on model performance.

Removing Color Features: We trained the model without color features to evaluate the reliance on shape information alone. As shown in Figure 3(a), performance did not improve, indicating that color features are not the sole cause of poor generalization.

Linear-Only Model: We tested a linear model without hidden layers to see if model complexity was contributing to overfitting. Figure 3(b) demonstrates that the linear model also failed to generalize, suggesting that the issue is not solely due to over-parameterization.

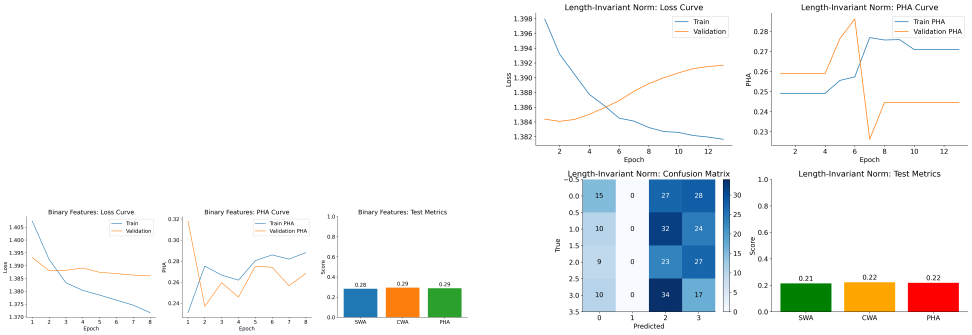


(a) Removing Color Features (b) Linear-Only Model

Figure 3: Ablation studies: (a) Removing color features did not improve performance, indicating that shape information alone is insufficient. (b) Using a linear-only model did not enhance generalization, suggesting that reducing model complexity does not address overfitting.

Binary Feature Representation: We experimented with binary feature representations instead of one-hot encoding. Figure 4(a) shows negligible performance changes, indicating that feature encoding choice is not a significant factor.

Length-Invariant Normalization: We applied length-invariant normalization to account for variable-length sequences. As depicted in Figure 4(b), this adjustment did not yield significant improvements.



(a) Binary Feature Representation (b) Length-Invariant Normalization

Figure 4: Further ablation studies: (a) Using binary feature representation did not affect performance significantly. (b) Applying length-invariant normalization showed negligible improvements, suggesting that sequence length variability is not the primary issue.

C Additional Confusion Matrix

To further illustrate the overfitting issue when training without early stopping, we present the confusion matrix in Figure 5. The confusion matrix shows increased misclassifications compared to models trained with early stopping, highlighting that extended training without proper regularization exacerbates overfitting.

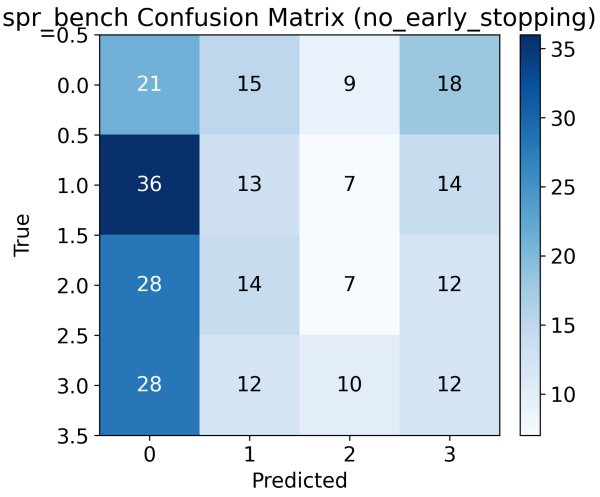


Figure 5: Confusion matrix for the model trained without early stopping. The model shows increased misclassifications, indicating severe overfitting due to the absence of early stopping.