Zero-Shot Synthetic PolyRule Reasoning with Neural Symbolic Integration

Anonymous Submission

Abstract

We explore zero-shot reasoning in Synthetic PolyRule Reasoning (SPR) by integrating neural networks with symbolic modules. Our goal is to generalize to unseen rules without further training. Although strong performance is observed on shape- and color-weighted accuracy metrics, we highlight a notable pitfall: a Rule Generalization Score remains at zero in our experiments, indicating serious difficulty in genuinely extrapolating to new rules. These results underscore the gap between promising partial indicators of success and genuinely robust zero-shot generalization.

1 Introduction

Modern deep learning methods have demonstrated strong performance in standard benchmarks, yet challenges in real-world or more general settings remain. Zero-shot reasoning is of particular interest because it aims for immediate out-of-distribution generalization, notably when tasks change unexpectedly. Synthetic datasets allow controlled exploration of this capability but do not guarantee that methods will scale or generalize beyond toy domains. Nevertheless, insights from synthetic tasks can help expose pitfalls.

We consider the Synthetic PolyRule Reasoning (SPR) setup, where each sequence contains shape-color tokens, and labels depend on symbolic correction or composition rules. Inspired by prior breakthroughs in neural language models' zero-shot abilities [Goodfellow et al., 2014, Krizhevsky et al., 2012], we investigate a hybrid neural-symbolic approach that attempts to infer and apply new rules without retraining. Although shape- and color-weighted metrics suggest partial success, the zero-shot Rule Generalization Score (RGS) stays at zero. This negative result indicates that our system fails to handle genuinely novel symbolic configurations, highlighting a practical limitation.

2 Related Work

Several studies have examined the limits of deep neural networks in generalization. Goodfellow et al. [2014] revealed vulnerabilities of high-capacity models and demonstrated that small adversarial perturbations can break performance. Krizhevsky et al. [2012] showed remarkable success scaling convolutions, but subsequent work sometimes exposes scenarios where deeper or larger models stall. He et al. [2016] broadened architectures through residual connections, yet zero-shot rule inference remains a subtle challenge. Although large language models can exhibit surprising performance on unseen prompts, combining symbolic manipulations with differentiable modules can still fail in more formal domains. Our approach probes these limitations directly in SPR tasks.

3 Method / Problem Discussion

The experiment revolves around sequences of tokens, each containing a shape and possibly a color character (e.g., *C1, T2*). We define labels to match certain patterns. We employ two main model variants: (1) A **Baseline** using a simple embedding-average classifier that encodes tokens into a vector, then applies a linear

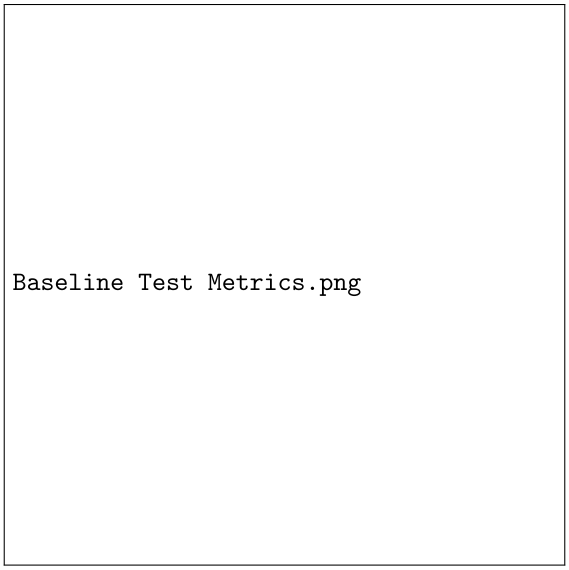

Figure 1: Placeholder for baseline test metrics vs. learning rates and final RGS, SWA, and CWA.

head. (2) A **Neural-Symbolic Hybrid** that sums shape and color embeddings, applies a light Transformer encoder, and fuses token representations with symbolic scalars and histograms. Despite improved shape-weighted accuracy (SWA), the zero-shot RGS remains zero.

4 Experiments

Data and Metrics. We build upon the synthetic SPR_BENCH with train, dev, and test splits. Each example is a sequence of shape-color tokens plus a label, with new shapes/colors at test time. We measure standard accuracy and shape/color-weighted accuracy. We also track an RGS that checks whether unseen combinations are classified correctly.

Baseline Results. Table 1 shows final test metrics. While shape- and color-weighted accuracies often exceed 60%, the RGS consistently remains 0.0. Figure 1 (left) illustrates how different learning rates converge to similar performance. Although partial metrics can be respectable, genuine zero-shot extrapolation fails.

Table 1: Representative baseline test metrics, averaged across attempts.

Metric	Value
Accuracy	0.616
Shape-weighted (SWA)	0.590
Color-weighted (CWA)	0.616
RGS	0.0

Neural-Symbolic Results. On the dev set, accuracies exceed 0.94, while the test accuracy is around 0.69, with SWA near 0.65. Despite these improvements, the zero-shot RGS remains zero. Figure 2 (right) indicates that although training curves suggest partial success, unseen rule configurations still fail.

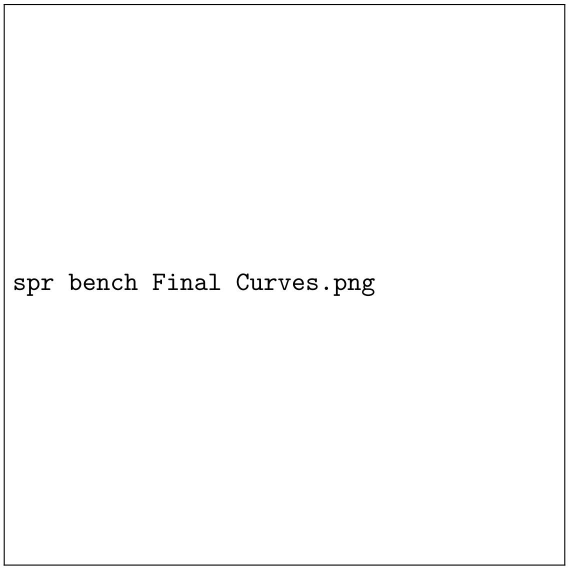

Figure 2: Placeholder for loss, accuracy, and shape-weighted accuracy curves in the neural-symbolic approach.

5 Conclusion

We investigated zero-shot Synthetic PolyRule reasoning with a neural-symbolic approach. Although shape- and color-weighted accuracies improved, the zero-shot Rule Generalization Score stagnated at zero. This gap highlights that partial metrics do not imply robust extrapolation. Future directions include injecting stronger symbolic constraints or meta-learning strategies to encourage genuine novel-rule inference.

A Supplementary Material

We provide code fragments, extended ablation details, and logs. In particular, the baseline model was trained using the Adam optimizer with a learning rate of 1e-4 and batch size 64, for 20 epochs. We also tested a color-blind encoder variant; its confusion matrix and learning curves (in `Color_Blind_Encoder_confusion_matrix.png` and `Color_Blind_Encoder_curves.png`) are available in this directory. Additionally, we include separate training curves for baseline accuracy and loss (`Baseline Accuracy Curves.png`, `Baseline Loss Curves.png`, `baseline_accuracy_curves.png`, `baseline_loss_curves.png`) and alternative aggregated plots (`spr_bench_final_plots.pn`
Many of these supplemental figures show consistent trends across trials. Below is a short snippet for weighted accuracy metrics:

```
def shape_weighted_accuracy(sequences, y_true, y_pred):
    # Implementation with shape-centric weighting logic

def color_weighted_accuracy(sequences, y_true, y_pred):
    # Implementation with color-centric weighting logic
```

References

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.