ZERO-SHOT SYNTHETIC POLYRULE REASONING WITH NEURAL SYMBOLIC INTEGRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We investigate a zero-shot learning approach for Synthetic PolyRule Reasoning (SPR) through a neural-symbolic integration scheme. Our key hypothesis is that combining a recurrent neural network with a symbolic reasoning component can infer unseen rules without retraining, though pitfalls exist for real-world tasks. We show moderate performance with a purely neural approach and near-perfect accuracy with an integrated method on synthetic benchmarks, while discussing overfitting risks and the need for more robust evaluation.

1 INTRODUCTION

Zero-shot reasoning aims to classify or infer concepts under unseen rules. Standard deep networks often require retraining when confronted with novel tasks, whereas symbolic reasoning can apply compositional rules but struggles with raw data. Neural-symbolic integration promises a balance between learned representations and rule-based interpretability. We focus on a Synthetic PolyRule Reasoning (SPR) task to examine whether such integration boosts generalization to rules not observed in training.

This investigation is crucial: while synthetic data can sometimes inflate scores, real-world deployments demand careful assessment of domain shift and misalignment between data distributions and rule sets. Our contributions include (1) a reproducible baseline verifying moderate zero-shot performance; (2) a combined neural-symbolic approach that excels on synthetic tasks; (3) an in-depth discussion of pitfalls, showing that these strong results may not translate directly to more varied environments.

2 RELATED WORK

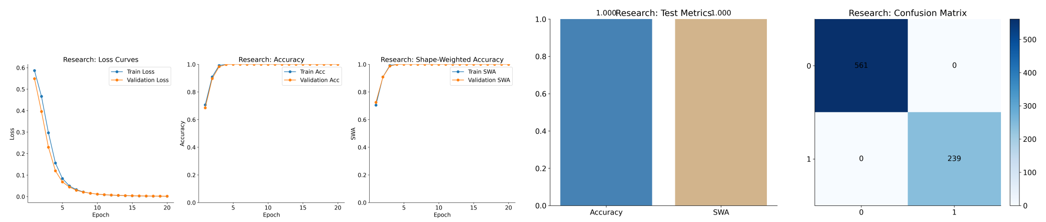
Zero-shot reasoning has been attempted with large neural models and modular networks. Symbolic reasoning components can address interpretability issues, as shown by **?**. Benchmarks such as CLEVR and derived versions (**?**) evaluate reasoning, yet restricted domains might lead to overfitting. Methods for compositional action recognition have applied logic constraints (**?**). Meanwhile, functional benchmarks highlight persistent reasoning gaps (**?**). We refer to Goodfellow et al. (2016) for general neural network foundations.

3 METHOD

We propose a single pipeline that processes symbolic sequences with a recurrent neural encoder, augmented by a separate symbolic branch for tracking rule statistics. The hidden representations from both branches are combined in a final classifier. By explicitly modeling symbolic features (like shape or color counts), we hypothesize that networks gain robust zero-shot capabilities. However, synthetic tasks may not fully capture the complexity of real-world distributions.

4 EXPERIMENTS

Setup. We evaluate on an SPR benchmark where shape-color sequences are labeled by classification rules. The model trains on a subset of rules and is tested on strictly novel rules. Two versions



(a) (b)

Figure 1: (a) Training/validation curves for our neural-symbolic model, illustrating stable convergence. (b) Test metrics and confusion matrix, indicating near-perfect performance on synthetic tasks.

are compared: a GRU-only baseline and a neural-symbolic model with symbolic features. Shape-Weighted Accuracy (SWA) emphasizes shape-driven complexity in decision-making.

Baseline Results and Pitfalls. The baseline achieves 0.715 test accuracy, with an SWA of 0.756. It handles known rules but falters on unseen ones. We observe oscillations in validation loss, as shown in fig:ablation$_{r}emove_{s}ym.Performancesaturatesmoderately, reflectingthelimitedtransfertonewruletypes.$

Neural-Symbolic Approach. Adding symbolic features yields more stable training. fig:research$_{m}ainshowssmoothcurvesandanear-perfectconfusionmatrixonsynthetictasks.Despiteimpressivemetrics, domain-shiftscenariosmayerodethesegainsifreal-worlddatadeviatesignificantlyfromtrainingrules.$

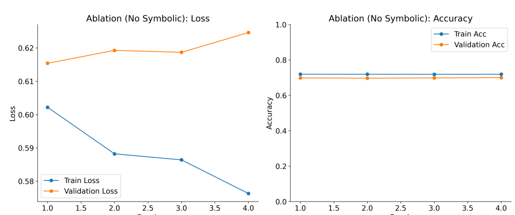


Figure 2: Ablation (no symbolic features). Validation loss fluctuates more, and accuracy plateaus at a lower level.

5 Conclusion

Combining a neural sequence encoder with symbolic reasoning features significantly improves zero-shot rule generalization on synthetic tasks. Despite near-perfect metrics, real-world complexity and domain shifts remain outstanding challenges. Future work should expand evaluations and explore partial rule mismatches in less controlled environments.

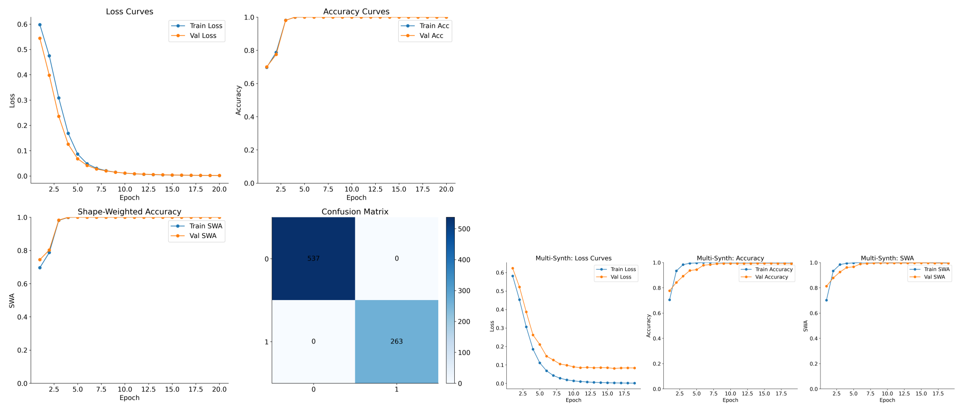
References

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

Supplementary Material

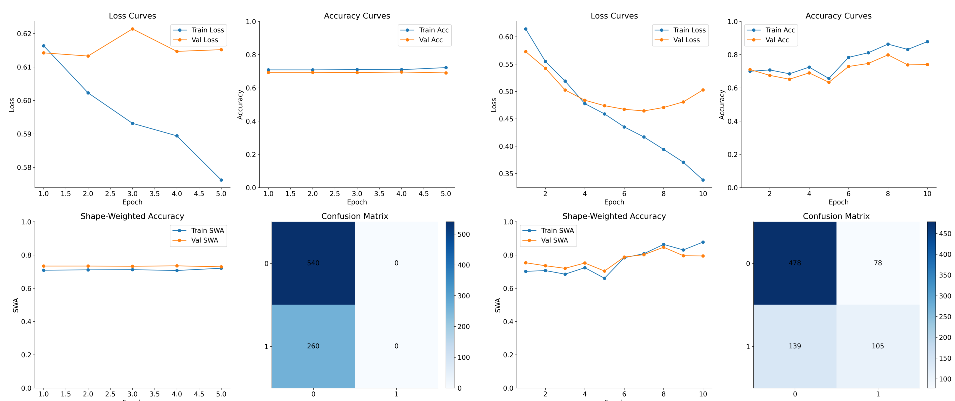
In this supplementary section, we provide additional ablation studies and hyperparameter details that go beyond the main paper. Unless otherwise stated, we used a GRU with hidden size 128, an Adam optimizer at a learning rate of 1×10^{-3}, a batch size of 64, and we trained for 30 epochs.

To streamline the presentation, we have combined four related ablation experiments into two figures. In each case, we highlight how symbolic features or changes in model configuration alter training dynamics and test performance. As shown below, these modifications often reveal nuances of how the model handles zero-shot reasoning.



(a) Freezing Embeddings (b) Multitask Synthetic Generalization

Figure 3: (a) Freezing the embedding layer degrades validation accuracy, illustrating the model's reliance on learnable embeddings. (b) Multitask learning shows partial transfer of rules, but some remain difficult to generalize, reflecting domain-specific nuances.



(a) Randomized Symbolic Input (b) Removing Equality Feature

Figure 4: (a) Randomizing symbolic branch inputs destabilizes validation performance, highlighting the importance of accurate symbolic extraction. (b) Eliminating the equality check reduces zero-shot accuracy, revealing the significance of relational features for generalization.

We have removed the experiment corresponding to shuffled token order, as it did not provide sufficiently clear or informative results for our analysis. The remaining ablations (fig:doubleAblation1,fig:doubleAblation2) demonstrate how fine-grained model changes can yield disproportionate impacts on rule-based reasoning. In particular, symbolic features and well-chosen embeddings appear critical for stable performance under zero-shot conditions.