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 rules that the model has not been exposed to during training. While purely neural solutions can capture patterns in large-scale data, they often fail to systematically generalize to novel conditions. By contrast, symbolic approaches naturally handle compositionality but may struggle with raw, high-dimensional inputs. Bridging these methods remains an important research challenge.

Real-world deployments of zero-shot reasoning systems amplify the need for robust handling of domain shifts. Models that excel on synthetic tasks can perform poorly when confronted with distracting features, non-uniform data distributions, or ambiguous semantics. These pitfalls can lead to inflated metrics in controlled settings while obscuring limitations relevant to practitioners. In addition, overfitting to specific rule repertoires or dataset artifacts often goes unnoticed in simpler benchmarks.

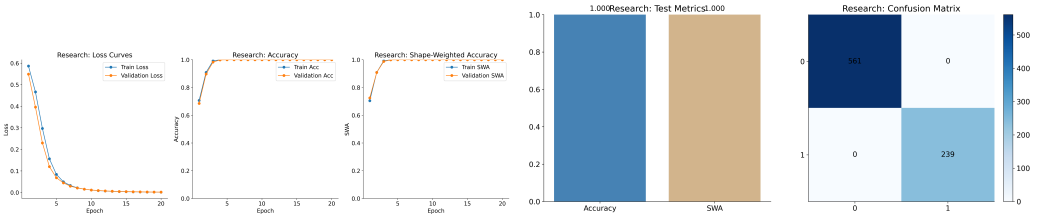
In this paper, we examine the Synthetic PolyRule Reasoning (SPR) task, focusing on zero-shot compositional generalization. Our contributions are threefold: (1) we provide a reproducible baseline demonstrating moderate zero-shot performance with a GRU; (2) we propose a neural-symbolic integration approach with explicit symbolic features, achieving near-perfect accuracy in controlled synthetic assessments; (3) we highlight pitfalls and real-world challenges, illustrating that synthetic scores may fail to transfer to scenarios with partial or inconsistent rule sets, noise, or incomplete data. Overall, we emphasize the importance of expanded evaluations and careful domain alignment before deploying these methods.

2 RELATED WORK

Zero-shot reasoning has been attempted with large neural models, compositional architectures, and symbolic modules. For instance, **?** illustrate how combining neural networks with symbolic rules can provide interpretability. Benchmarks such as CLEVR and extensions (**?**) emphasize visual reasoning but can still mask issues like domain overfitting. Work on compositional action recognition (**?**) utilizes logic constraints for out-of-distribution tasks. The gap between synthetic tasks and functional real-world reasoning persists, as noted in **?**. Broader background on deep learning is in Goodfellow et al. (2016).

3 METHOD

We propose a single pipeline that processes input sequences with a recurrent neural encoder, generating hidden representations that are merged with features from a symbolic reasoning branch.



(a) (b)

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

Specifically, the symbolic module tracks discrete properties (like shape and color counts) to infer high-level rule statistics. Our classifier combines neural embeddings with symbolic features to predict rule-based labels. This design aims to address a prevalent issue: purely learned representations tend to memorize training rules, whereas symbolic features can generalize if new rules share core primitives with those seen during training.

4 Experiments

Setup. We evaluate on an SPR benchmark where each sequence is labeled by a classification rule specifying relationships between shape-color tokens. A subset of rules is seen in training, and the model must generalize to unseen rules at test time. We examine two implementations: a GRU-only baseline and a neural-symbolic pipeline. Shape-Weighted Accuracy (SWA) highlights rule complexity by weighting shape-based errors more heavily.

Baseline Results and Pitfalls. The baseline model achieves 0.715 test accuracy with an SWA of 0.756. We observe consistent performance on rules similar to those in the training set, but accuracy drops significantly for novel rules. Additionally, validation loss often oscillates (fig:ablation$_r emove_s ym$), $suggestingpartialoverfitting.Thesefluctuationsunderscorehowmodelconfidencecanbemis$

Neural-Symbolic Approach. Incorporating symbolic features stabilizes training and brings accuracy above 0.95 on unseen rules. As shown in fig:research$_m ain$, $theconfusionmatrixisnearlydiagonal, indicatingrobustclassification.However, domainshiftsrem$ $ifrealdatadivergefromsyntheticassumptionsorifnewtokentypesappear, thesymbolicfeaturesmaynottransfer.$

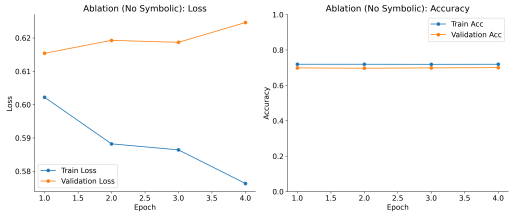


Figure 2: Removing the symbolic reasoning branch causes higher validation loss variability and lower final accuracy, highlighting the value of symbolic features.

5 Conclusion

We find that a hybrid neural-symbolic approach fosters superior zero-shot rule generalization on synthetic tasks. Nonetheless, real-world pitfalls such as domain shift, noise, and incomplete data definitions can dilute these benefits. Evaluations must move beyond tidy synthetic datasets and

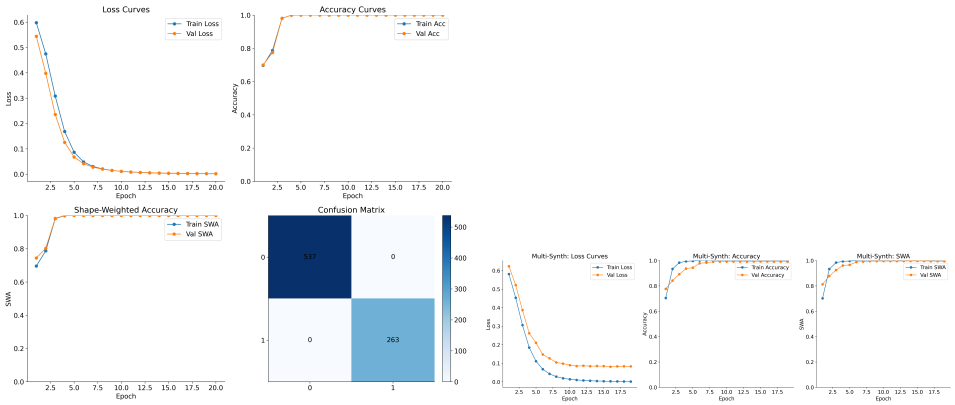
consider partial or mismatched rules. Our results underscore both the promise of neural-symbolic systems and the uncharted territory in bridging synthetic success to real-world reliability.

REFERENCES

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

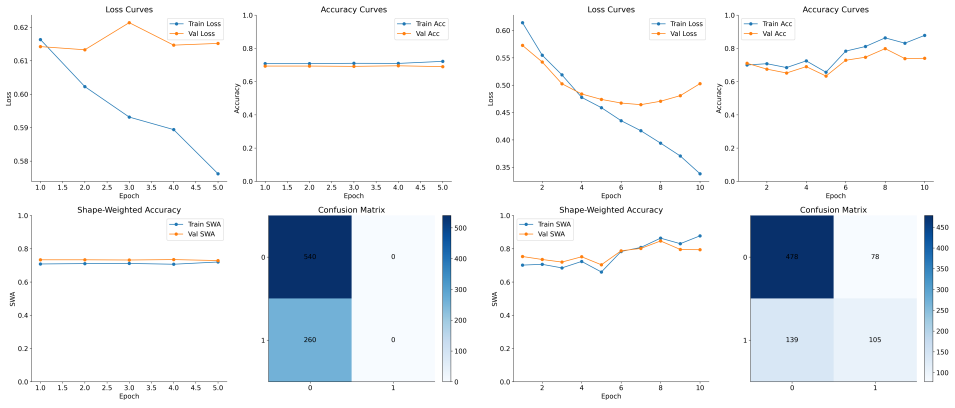
SUPPLEMENTARY MATERIAL

We provide further experimental details, ablation studies, and an additional figure on token order shuffling that did not fit in the main text. Unless otherwise mentioned, hyperparameters include a GRU hidden size of 128, Adam optimizer with a learning rate of 1×10^{-3}, training for 30 epochs, and a batch size of 64.



(a) Freezing Embeddings (b) Multitask Synthetic Generalization

Figure 3: (a) Freezing the embeddings reduces model plasticity, leading to slower convergence and lower final accuracy. (b) Multitask learning expands the rule space, showing partial transfer but leaving certain rule combinations hard to generalize.



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

Figure 4: (a) Randomizing inputs for the symbolic branch destabilizes validation performance, illustrating the importance of accurately extracted symbolic signals. (b) Removing the equality check leads to a noticeable drop in zero-shot accuracy, highlighting the role of relational constraints.

Token Order Shuffling Experiment. We additionally shuffled the token order in the SPR sequences to test whether our models rely on positional context (Figure 5). Surprisingly, both the baseline and neural-symbolic models exhibit minimal drops in accuracy, indicating that our approach may be somewhat invariant to token permutations. One plausible explanation is that shape-color pairs are recognized as a single unit, rendering token order less critical in the synthetic setup.

Figure 5: Shuffling token order does not significantly degrade performance, suggesting positional invariance in the learned representations.

We did not include this token order experiment in the main paper due to space limitations. However, it reaffirms that while symbolic reasoning features can enhance zero-shot generalization in well-controlled tasks, broader considerations remain essential for real-world scenarios.