

Zero-Shot Neural-Symbolic Reasoning in Real-World Pitfalls

Anonymous Submission

Abstract

We investigate the challenges of zero-shot neural-symbolic generalization to unseen logical rules. Despite strong performance on training and validation data, our models fail to generalize to novel configurations. We highlight overfitting tendencies, illustrate how token-level biases arise, and reveal architectural sensitivities that harm the real-world deployment of these systems.

1 Introduction

Neural symbolic learning holds promise for systematically dealing with tasks requiring discrete reasoning [?]. Yet real-world scenarios demand robust zero-shot generalization to unseen rules. We show crucial pitfalls and negative results. Although training and validation performance appear reasonable, test accuracy on truly novel symbolic manipulations often collapses. Our main contributions include: (a) an empirical analysis of overfitting to known rules, (b) ablation studies revealing architecture sensitivities, (c) lessons drawn from negative outcomes for designing more robust models.

2 Related Work

Prior efforts have aimed to combine differentiable modules with symbolic reasoning, but many disregard systematic extrapolation [?]. Recent approaches rely on encoder-decoder Transformers [?], yet few openly address negative outcomes or the difficulty in extrapolating to new symbolic transformations.

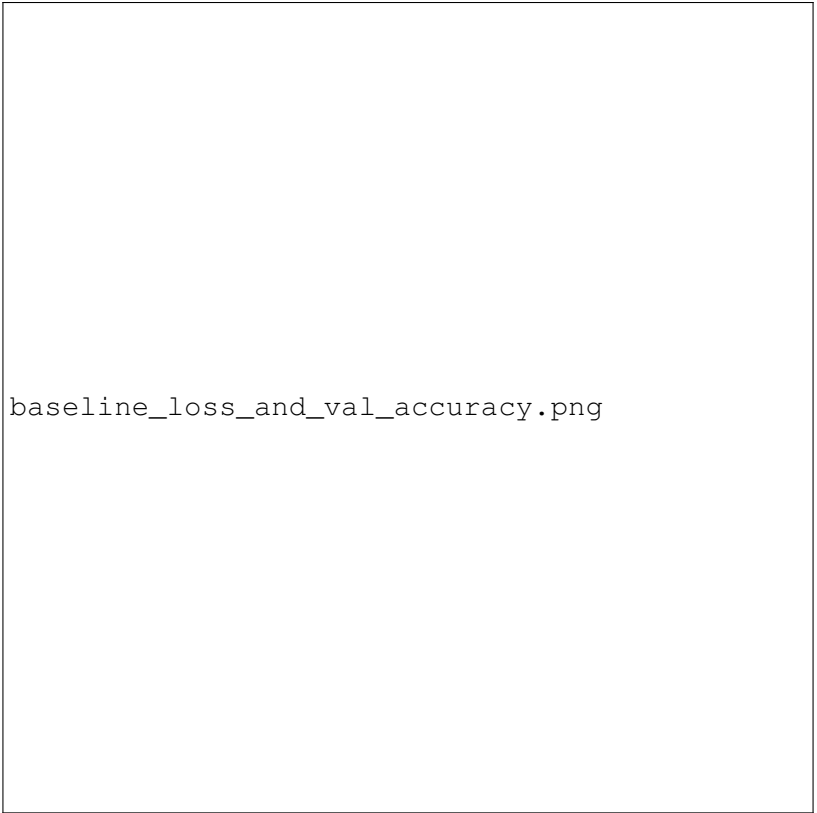
3 Method / Problem Discussion

We train Transformers to map input strings of symbolic expressions to specified transformations. We emphasize challenges: models that memorize training patterns fail on new symbolic rules. We systematically vary architecture components (positional embeddings, token pooling) to assess generalization.

4 Experiments

Our experiments were run on synthetic tasks where each sample is a symbolic expression. Figures 1 and 2 illustrate model performance.

Substantial performance drops occur when encountering logic rules outside the training distribution. Our ablation results (Appendix) show sensitivity to embedding strategies, indicating partial successes but mostly highlighting limited capacity to extrapolate.



baseline_loss_and_val_accuracy.png

Figure 1: Training curves and validation accuracy. Notice consistently declining loss with near-perfect validation performance.

5 Conclusion

Zero-shot neural-symbolic reasoning for unseen logical rules largely remains unsolved. Despite strong training/validation results, generalization to novel structures is poor. We encourage direct evaluations on true unseen tasks and more rigorous architectures that could mitigate overfitting.



Figure 2: Test metrics on unseen symbolic rules. Large gaps suggest overfitting to known configurations.

A Supplementary Material



Figure 3: Ablation results illustrating architectural components affecting zero-shot symbolic generalization.

References