Exploring Unmet Challenges in Symbolic Reasoning with Neural Networks

Anonymous Submission

Abstract

Although deep learning has advanced many fields, certain tasks still exhibit performance plateaus. We investigate a symbolic reasoning task where state-of-the-art models yield only partial success, highlighting real-world pitfalls such as overfitting and brittleness. Our findings emphasize that blindly scaling models may not solve fundamentally structured tasks and could hinder reliable deployment.

1 Introduction

Neural networks excel at tasks with large labeled datasets [?]. However, tasks demanding robust symbolic reasoning remain challenging. We present a systematic study revealing overfitting, minimal gains from increased model size, and inconsistent results across seeds. Our contributions include negative results on scaling methods and a clearer understanding of where typical optimization schemes [?] fall short. These insights can serve as cautionary tales for practitioners.

2 Related Work

Prior research on language modeling [?] and domain-specific reasoning has shown impressive gains but also uncovers domains resistant to purely data-driven approaches. Some highlight that models may overfit or struggle with discrete structures. Our findings corroborate these pitfalls, emphasizing the importance of carefully interpreting partial improvements and inconclusive results.

3 Method

We employ a standard encoder-decoder architecture to tackle symbolic deductions. Despite extensive hyperparameter tuning (details in Appendix), models often hit a performance ceiling. We hypothesize that data augmentation or specialized architectures might be necessary, but our attempts yielded only marginal gains.

4 Experiments

We compare a baseline and a research variant on a novel symbolic reasoning dataset. Training/validation curves in Figure 1 show similar learning progress. The final evaluation (Figure 2) reveals that scaling parameters or training steps yields diminishing returns. Divergences among random seeds further emphasize the fragility of these models.

5 Conclusion

We show that popular neural architectures may plateau on symbolic tasks, leaving substantial gaps. Our investigation underscores the need for novel approaches that combine structure-aware methods with standard deep learning. Future work might integrate modular components or data-centric strategies to overcome these limitations.

References

Figure 1: Aggregated macro-F1 and loss trends. Both models plateau early, suggesting limited benefit from further training.

Figure 2: Test macro-F1 comparison between baseline and proposed approach. Gains remain modest.

A Appendix

Additional results and hyperparameters are provided here. Supplementary epoch-level comparisons, merged confusion matrices, and per-seed analyses are included for completeness.

baseline_test_macro_f1_vs_epochs.png

research_avg_loss.png

research_confusion_matrices.png

research_confusion_matrix_detail.png