Revisiting Instabilities in Symbolic Reasoning: A Cautionary Tale

Anonymous Submission

Abstract

We investigate challenges associated with applying deep neural networks to symbolic reasoning tasks. Our experiments reveal instabilities and partial improvements that fail to generalize. These issues highlight key pitfalls in real-world deployment.

1 Introduction

Despite the continued success of deep learning in various tasks, symbolic reasoning remains a notoriously difficult application. Previous work has shown that even seemingly straightforward symbolic constraints can lead to unexpected behaviors when integrated into neural models [**?**, **?**]. Here, we present experiments that corroborate and extend these observations. We find that subtle changes in data distribution or hyperparameters can produce significant performance swings, underscoring the difficulty of reliably handling symbolic reasoning in real-world contexts [**?**].

reflection$_p age_i nfo$

2 Related Work

Symbolic tasks have been explored extensively, yet existing methods often overlook the brittleness of learned representations [**?**]. Our focus is on highlighting these pitfalls and partial progress. Recent studies analyze how architectures fail under constrained logic, pointing to a need for deeper insights [**?**, **?**].

3 Method Discussion

We employ a series of neural baselines (GRU, Transformer) for a symbolic reasoning benchmark. Models are trained to predict consistent labels under compositional constraints. In our setup, minor changes in the data or training routine often lead to erratic behaviors and inconclusive performance gains.

4 Experiments

We summarize two primary experiments, providing insight into the pitfalls:

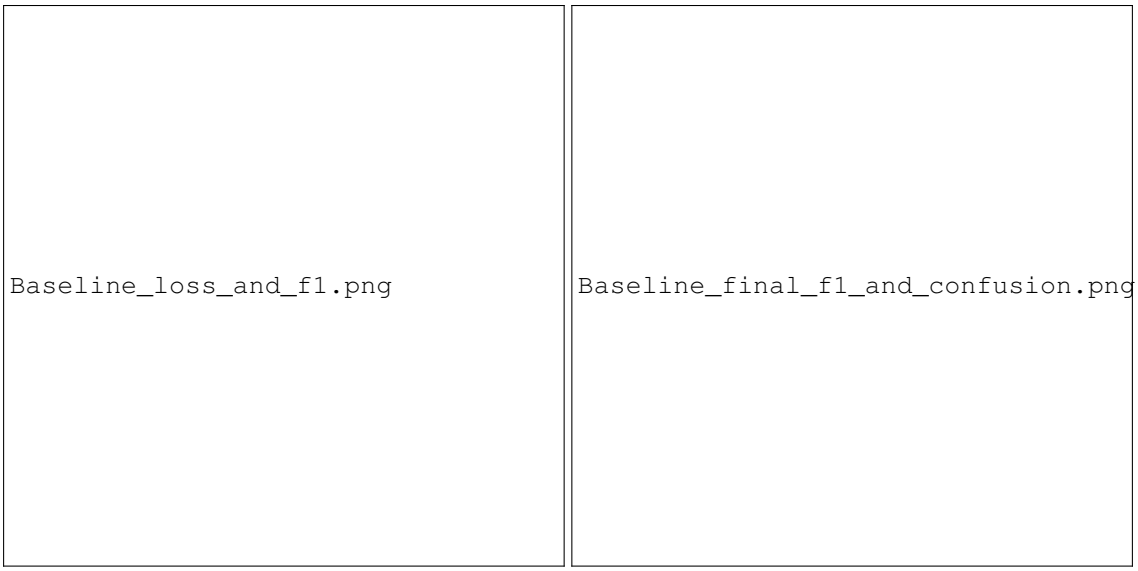

Figure 1 illustrates training instability and confusion matrix outcomes for a GRU-based baseline.

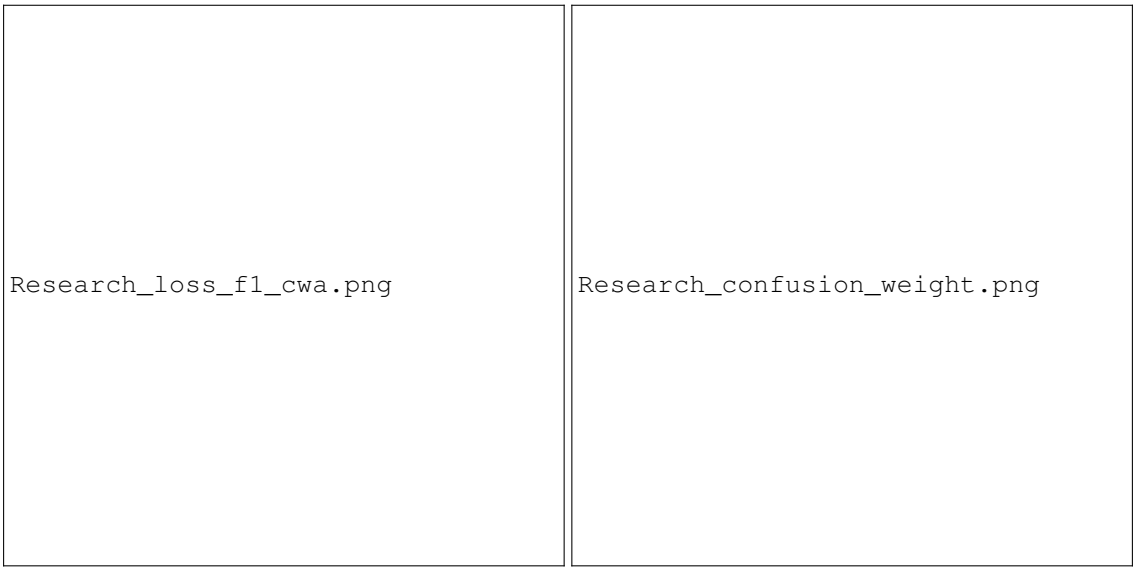

Figure 2 highlights the complexity-weighted accuracy of a Transformer model, showing partial gains yet susceptibility to certain classes.

5 Conclusion

We show that standard deep models can exhibit unstable or incomplete solutions to symbolic tasks. Lessons learned indicate that careful data curation, hyperparameter tuning, and interpretability measures are critical. Future work lies in designing robust interventions to handle compositional constraints more reliably.

References

A Additional Experiments

Further ablation results and training configurations are presented here. We include additional confusion matrices, label distributions, and performance metrics that validate our main claims while illustrating overlooked complexities in symbolic reasoning.