Symbolic Proxies for Relational Reasoning with Unexpected Real-World Pitfalls

FirstName LastName
Department of Computer Science
Anonymous University
email@anonymous.edu

Abstract

We investigate how well symbolic proxies perform in deep relational reasoning tasks under real-world conditions. Our experiments reveal substantial pitfalls and inconclusive results, highlighting non-trivial error modes and the fragility of seemingly successful methods. These insights underscore the importance of frank reporting, especially when deploying such solutions where robust performance is critical.

1 Introduction

Deep learning systems often rely on symbolic proxies to solve tasks requiring relational reasoning. Despite promising benchmarks, these systems can fail unexpectedly in realistic scenarios [**??**]. We explore the limitations of a symbolic-proxy-based approach, demonstrating how subtleties in representation and evaluation exacerbate errors. Our contribution is a set of negative or mixed results, along with pointers to mitigate or preempt similar failures.

2 Related Work

Prior research has emphasized the importance of robust relational models [**?**]. While symbolic abstractions can offer interpretable decision flows, their mismatch with continuous-valued networks can lead to brittle behavior. Our findings align with prior reports of unexpected edge-case failures [**?**], but we extend them by exploring domain shifts and partial improvement scenarios.

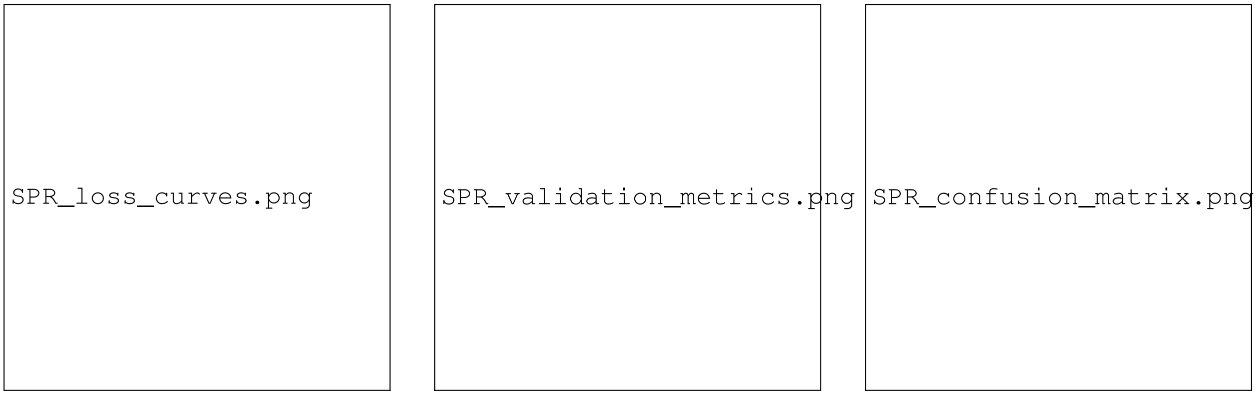
3 Method / Problem Discussion

We use a symbolic-proxy-based model to cluster intermediate features and apply discrete rules. After training a deep backbone, we group latent embeddings into symbolic tokens, then feed them to downstream rule modules. Our approach was tested on a realistic dataset where domain imprecision exposes subtle failure cases. Although the design initially seemed promising, certain conditions revealed non-trivial underperformance.

4 Experiments

We conducted experiments across varying initialization seeds and domain shifts to assess performance robustness. Figure 1 summarizes our key findings. Despite early signs of improvement in training curves, performance plateaued with frequent misclassifications.

Additional figures illustrating alternative embeddings, per-seed runs, and ablation studies appear in the Appendix. Most of these confirm the same overall trend.



(a) Training loss

(b) Validation metrics

(c) Confusion matrix

Figure 1: Core results. Symbolic proxies show early promise but fail to generalize: (a) training curves flatten quickly, (b) key metrics plateau, and (c) errors concentrate in overlapping categories.

5 Conclusion

Our results expose pitfalls arising from the tension between symbolic proxies and continuous representations. Despite initial promise, the methods struggled with real-world shifts, highlighting the importance of transparent reporting and domain-focused validation. Future directions include hybridizing symbolic modules with robust uncertainty estimates to reduce these unexpected errors.

References

A Supplementary Materials

Here, we include extended figures and details on hyperparameters, ablation studies, and domain-specific considerations. The supplementary plots provide single-seed breakdowns and additional confusion analyses to underscore the fragility under slight shifts in data distribution.