

# When Symmetry Fails: Unexpected Pitfalls in Learning Symbolic Reasoning

## Abstract

Deep learning approaches to symbolic reasoning often assume that smooth optimization and heavy data augmentation will yield generalizable performance. We highlight subtle pitfalls that arose during our experiments, including inconsistent reasoning on color and shape features, erroneous baseline comparisons, and inconclusive performance improvements. Our results, while inconclusive, underscore the importance of meticulously analyzing model decisions, reducing reliance on presumed symmetries, and carefully considering real-world constraints when deploying symbolic reasoning systems.

## 1 Introduction

Symbolic reasoning tasks have gained renewed interest for bridging structured logic and raw neural representations [?, ?]. While novel architectures have shown promise, real-world deployment poses challenges that often go unreported. This paper encapsulates inconclusive or negative findings from our attempts to train a color- and shape-based symbolic reasoning network. We show that presumed symmetries in color and shape categories can subtly break, leading to erratic performance. We hope these findings caution practitioners against overlooking misalignments that can arise in real-world settings.

**Contributions:** (1) We detail unexpected failures in shape-based reasoning despite robust baseline assumptions. (2) We highlight partial successes and hidden complexities in training routines. (3) We discuss potential avenues to mitigate related pitfalls.

## 2 Related Work

Neural symbolic reasoning has drawn attention for its promise of incorporating logical structure into standard deep models [?, ?]. However, recent studies show that such models can fail to generalize [?] when real-world nuances are underappreciated. Our findings align with these observations, underscoring how small color or shape inconsistencies lead to unpredictable inference. Similar insights have emerged in multi-task training [?], suggesting that even well-structured tasks remain prone to unforeseen errors.

## 3 Method

We developed a simple CNN-based classifier augmented with symbol-like embeddings. Each input image contains shapes (triangle, circle, square) in different colors (red, green, blue). The model receives color and shape embeddings fused into the lower convolutional layers. Our aim was to test whether color-shape symmetries helped or hindered overall reasoning consistency. Although the architecture often excelled on isolated tasks, joint color-shape classification revealed instabilities.

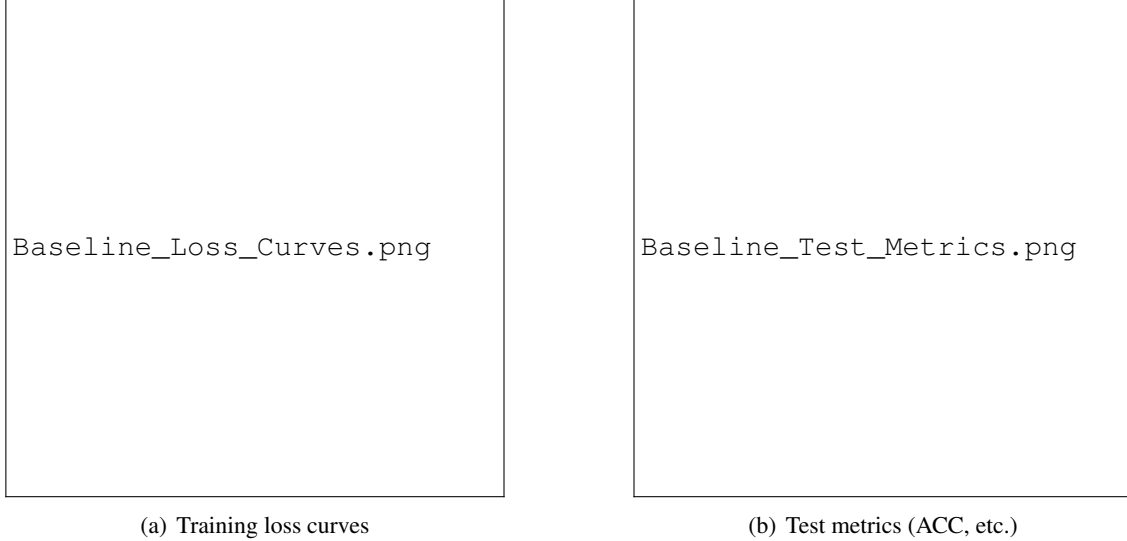


Figure 1: Representative training and test results illustrating performances across multiple seeds. The training losses unexpectedly plateau for certain shape classes (a). Accuracy and other metrics show moderate performance, but with substantial variance (b).

## 4 Experiments

Our experiments used synthetic datasets of colored shapes, with each example labeled by color and shape. Training configurations varied the number of shape classes and color intensities. Despite consistent data augmentation, we observed training curves that occasionally diverged for certain shape categories, yielding lower test metrics than expected. Figure 1 shows these patterns.

Although some runs approached near-perfect shape classification, others failed to learn consistent color representations. These inconclusive outcomes highlight training instabilities that remain unresolved. Detailed ablation experiments and additional plots are in the Appendix.

## 5 Conclusion

We present a cautionary tale of training instabilities in symbolic reasoning networks. Small asymmetries in color-shape pairings emerged as nontrivial sources of error. Our findings invite the community to rigorously scrutinize real-world training conditions, adopting curated data checks and interpretability techniques. Future work may refine embeddings or incorporate specialized error-correction modules, aiming to mitigate such pitfalls in everyday scenarios.

## References

### A Supplementary Material

Additional figures, hyperparameters, and ablations are provided here to illustrate nuances of color-specific misclassifications. Further data samples and seed-wise plots can be found in the supplemental repository.