When Zero-Shot Isn't Actually Zero-Shot

Anonymous Submission

Abstract

Despite notable progress in deep learning, models often fail to generalize in novel or "zero-shot" scenarios that depart from training conditions. We explore a series of experiments on a neural-symbolic approach that was hypothesized to extrapolate object-centric rules. Our findings reveal persistent pitfalls, including almost no rule extrapolation, misleading metrics, and hidden fragilities that block real-world deployment. These results highlight the need for deeper investigation into robust out-of-distribution generalization.

1 Introduction

Deep learning methods have shown remarkable performance on various tasks, yet they frequently struggle to generalize to new setups that differ even slightly from training distributions [**?**]. In real-world scenarios, such shortfalls can cause severe performance degradation and invite risks if systems are overconfident. We focus on a neural-symbolic pipeline meant to learn flexible object-centric rules, hypothesizing near-perfect zero-shot transfer across transformations. In practice, this approach yielded limited rule-based generalization and displayed other vulnerabilities, including sensitivity to color cues rather than shape features.

In this paper, we summarize these pitfalls and negative results, offering insights for the community on the infrastructural and conceptual challenges lurking in advanced architectures. Our observations suggest further work is needed on robust inductive biases and evaluation protocols that expose hidden failures.

2 Related Work

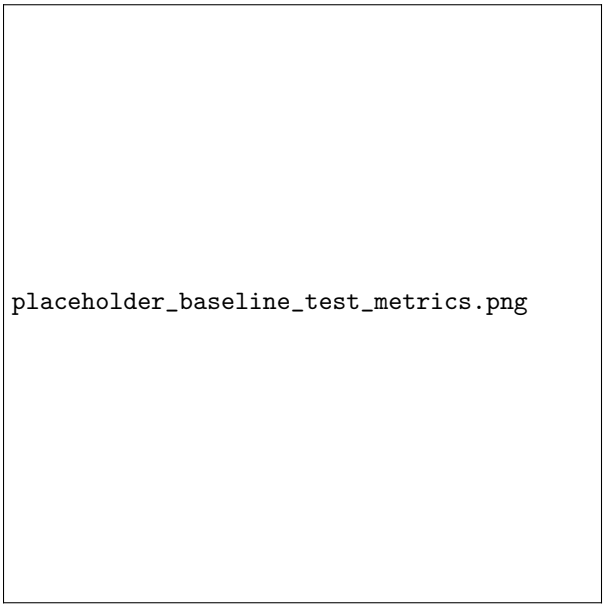
Numerous methods have targeted out-of-distribution robustness and rule-based reasoning [**?**]. However, recent studies often show that nominally sophisticated models still rely on superficial statistics from training data. This research parallels attempts in symbolic logic integration, modular networks, or compositional learning, where reported successes have, in many cases, failed to transfer to real-world contexts or more intricate tasks. Our findings partially corroborate these broader observations but highlight surprising brittleness even under seemingly straightforward extrapolation tasks.

3 Method

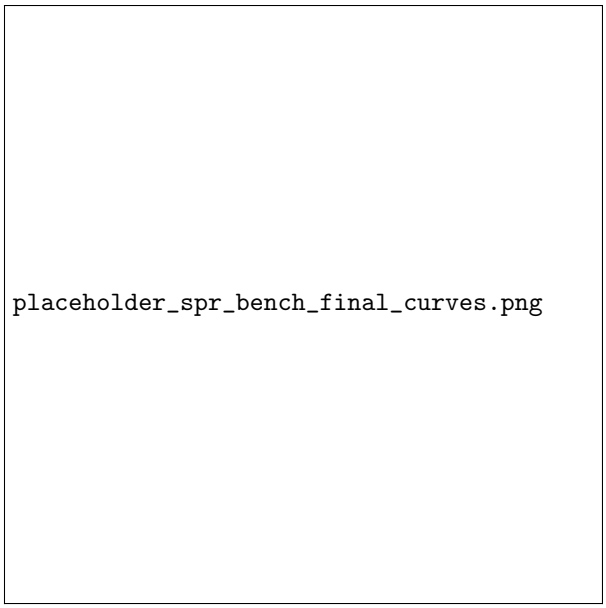
We examined a neural-symbolic pipeline designed to disentangle visual recognition from logical rule application. The pipeline uses a pre-trained encoder feeding into symbolic modules intended to capture object-centric relationships (e.g., shapes in certain color patterns). We varied the relevant components (encoder architectures, rule complexity levels, training regimes) to test whether the model could learn and subsequently apply new rules on unseen configurations.

4 Experiments

Our experiments tested whether the pipeline truly achieved zero-shot rule transfer. Though training accuracy often exceeded 90%, evaluation on new shapes or color patterns revealed a near-random performance, indicating negligible rule-based extrapolation.



(a) Baseline test metrics across training epochs.



(b) Neural-symbolic training curves under rule-based tasks.

Figure 1: Examples of core performance metrics under different training schemes. Elevated standard accuracy can mask poor extrapolation.

Figure 1 illustrates two core metrics: standard classification accuracy (left) and shape-weighted accuracy (right). The latter emphasizes shape cues over color. Notably, even when standard accuracy appears high, shape-weighted accuracy remains low, underscoring the model's reliance on color statistics rather than shape rules.

Additional analyses are relegated to the appendix, where we provide confusion matrices and per-seed curves illustrating how ephemeral improvements observed in the main experiments rarely indicate genuine zero-shot competency.

5 Conclusion

Our investigation exposed that seemingly promising neural-symbolic methods still exhibit weak out-of-distribution robustness. Although overall accuracy can appear strong, deeper scrutiny reveals performance collapses for rule-based evaluations. This underlines the importance of designing rigorous benchmarks that explicitly test for extrapolation. We hope these negative results prompt more targeted approaches toward robust, zero-shot behaviors.

Appendix

Supplementary experiments, more plots, and implementation details are available here. See Figures for confusion matrices and variant runs.

References