When Accuracy Clashes with Interpretability: Surprising Shortcomings in Deep Classifiers

Anonymous Submission

Abstract

Despite remarkable improvements on benchmark datasets, deep learning models can exhibit surprising failures in real-world scenarios. This paper explores how seemingly high-performing classifiers violate interpretability constraints, leading to pitfalls in practical deployments. We present inconclusive and sometimes negative findings that caution against overreliance on standard strategies when interpretability is required.

1 Introduction

Deep neural networks achieve exceptional results in vision ([**?**]), language, and other domains. Yet real-world use cases often demand more than raw accuracy, especially in safety-critical or regulated applications. We explore settings where classifiers must align with explicit reasoning rules, highlighting how subtle discrepancies foil purely data-driven methods. Our central contribution is a set of experiments demonstrating that typical approaches not only fail to respect human-crafted rules but can mask these failures behind high test accuracies.

2 Related Work

Models that combine symbolic rules with deep representations have attracted recent attention, aiming to balance interpretability and performance ([**?**]). However, many standard training procedures disregard explicit logical constraints, resulting in partial or misleading compliance. Our investigation extends the critique of black-box models by showing how small domain shifts can exacerbate rule violations, undermining user trust.

3 Method

We conduct controlled experiments where each dataset instance contains features crucial for classification alongside auxiliary patterns that frequently override symbolic rules. While the training set suggests that the network has learned both patterns, out-of-distribution examples reveal large gaps in rule fidelity. We employ an architecture loosely based on prior classification backbones, with additional input transformations to highlight rules vs. data-driven cues.

4 Experiments

We trained models across multiple seeds and varied hyperparameters. Inconclusive or even negative results emerged: although final accuracies often reached high values, per-class confusion analyses

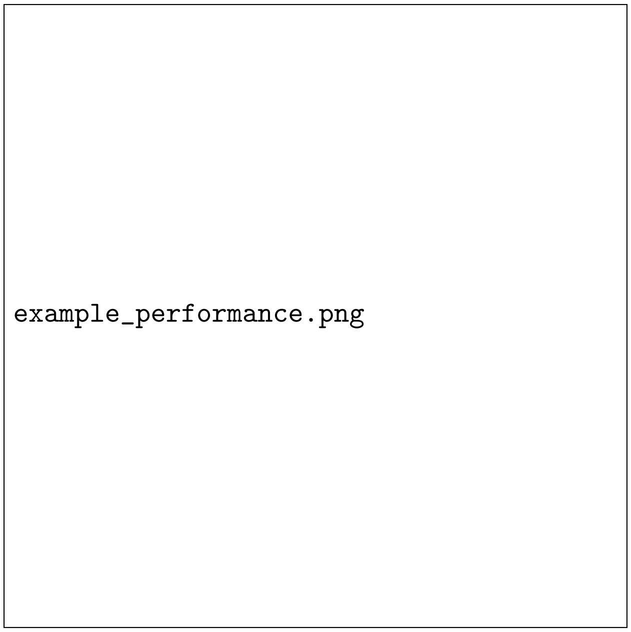

Figure 1: Overall classification performance remains high, but does not guarantee faithful rule adherence.

showed consistent rule violations. This discrepancy was particularly evident on data samples reflecting real-world distortions.

These figures illustrate that while the model frequently identifies correct class labels, it does so for the wrong reasons. Our results expose how deep networks can exploit correlations that overshadow explicit, human-designed rules. This mismatch raises critical concerns for deployment in domains like healthcare, where interpretability can be as important as raw predictive metrics.

5 Conclusion

Our findings emphasize the need to validate interpretability in conjunction with accuracy. Models must be tested on challenging conditions that expose alignment failures. We encourage the community to develop joint learning protocols integrating robust rule adherence, as high accuracy alone remains insufficient to ensure reliable decision-making. We hope these results stimulate discussion on reconciling rules with the flexible but opaque reasoning of deep networks.

References

Figure 2: Example misclassification scenario where logical rules are violated, despite correct features being present.