Unexpected Pitfalls and Partial Insights in Exploring Model Interpretability

Anonymous Submission

Abstract

We examine a series of negative or inconclusive findings associated with model interpretability in deep learning systems deployed in real-world contexts. Our work highlights subtle pitfalls that hinder the reliable interpretation of model outputs, undermining trust in critical applications. Understanding these challenges is a vital step toward building more robust explanations in practice.

1 Introduction

The deployment of deep learning systems in high-stakes domains has increased the demand for model interpretability. Although early research suggested promising directions for generating explanations, many of these methods face challenges when applied under realistic conditions. Our primary motivation is to report practical pitfalls observed during attempts to interpret deep models in industrial settings.

We conduct a systematic exploration of methods widely touted for their interpretability. Rather than achieving a definitive improvement, we encounter hurdles that cause discrepancies between theoretical expectations and real-world outcomes. Our contributions highlight these inconclusive results and discuss why even state-of-the-art interpretability techniques might lead practitioners astray.

2 Related Work

Several approaches have been suggested to uncover insights in deep models. Saliency maps and related methods are often used to pinpoint regions in an input image or features from tabular data that contributed strongly to a model's prediction [**?**]. More advanced frameworks, such as those based on deep residual networks, provide suggestive visual explanations without guaranteeing robustness [**?**]. Comparative analyses frequently focus on the success cases, leaving negative or ambiguous findings underreported. This paper aims at filling that gap.

3 Method Discussion

We focus on interpretability tools applied to multiple architectures ranging from convolutional to attention-based models. Our goal is not to propose an alternative explanation algorithm, but rather to test how effectively existing methods translate to actual usage. Our experiments revolve around standard datasets and some custom data from real-world industrial pipelines. Despite carefully following procedure for each method, we observe behavior that raises concerns about trust and reproducibility.

For instance, certain sets of features that appear salient in one context fail to generalize when small perturbations are introduced. Even more concerning, we find that repeated runs on identical data produce substantially different interpretations, bringing into question the stability of these insights.

4 Experiments

We evaluate the interpretability methods under conditions designed to model practical challenges: noisy data, distribution shifts, and limited annotation quality. Numerical metrics, alongside qualitative assessments, illustrate the inconsistent reliability of saliency-based approaches. Although some techniques appear robust initially, their performance declines sharply when confronted with real-world data complexities.

Tables summarizing accuracy or agreement scores provide further evidence of these pitfalls. In several cases, interpretability metrics show contradictory patterns across different seeds, emphasizing that random initialization can distort the perceived importance of certain features. Despite repeated fine-tuning and hyperparameter adjustments, consistent improvements were elusive, underscoring the difficulty of interpretability in real-world deployments.

5 Conclusion

We have highlighted a series of problems encountered when applying interpretability approaches to modern deep learning systems in realistic contexts. Our negative and inconclusive results raise awareness about the limited reliability of commonly adopted methods. Valuable future directions include examining how stable attributions can be guaranteed, whether through improved theoretical grounding or standardized evaluation protocols that address noise and distribution shift. Our experiences suggest that building trustworthy explanations for deep models remains a challenge requiring deeper investigation by the research community.

Appendix

Further experimental details, including hyperparameters, extended tables, and plots are presented here. No new figures are included in the main text due to limited added value for our conclusions. We encourage readers to verify the aggregated data and supplemental results to confirm the reproducibility of our negative findings.

References