Certifying Differential Invariants of Neural Networks using Abstract Duals

CHANDRA KANTH NAGESH, University of Colorado Boulder, USA

Neural networks are increasingly used in safety-critical domains, necessitating formal guarantees about their properties under perturbations to inputs. Existing robust verification techniques, typified by `DeepPoly`, primarily focus on the Forward Pass Verification Problem—certifying output stability where $f(\mathbf{I}) \subseteq [\mathbf{y}_L, \mathbf{y}_R]$ for an $\mathbf{L}_\infty$-constrained input set $\mathbf{I} = \{\mathbf{x} \mid \|\mathbf{x} - \mathbf{x}_0\|_\infty \leq \epsilon\}$. While methods like `DeepPoly` employ a sophisticated polyhedral abstract domain (combining intervals and affine forms) to generate sound, tight bounds, this entire class of analysis still fails to provide sound guarantees over the network's derivative behavior. This oversight creates a critical verification gap for gradient-dependent systems—such as scientific machine learning models and feedback control loops—where the core challenge shifts to the Backward Pass Verification Problem: formally bounding the Jacobian $\mathbf{J}(\mathbf{x}) = \nabla_\mathbf{x} f(\mathbf{x})$ such that $\mathbf{J}(\mathbf{x}) \subseteq [\mathbf{J}_L, \mathbf{J}_R]$ for all $\mathbf{x} \in \mathbf{I}$. This new capability is essential for certifying crucial differential invariants.

1 Introduction

2 Overview

3 [Contribution 1]

4 [Contribution 2]

5 Evaluation

6 Related Work

7 Conclusion

Acknowledgments

TBD

References

Author's Contact Information: Chandra Kanth Nagesh, ckn@colorado.edu, University of Colorado Boulder, USA.